DOCUMENT RESUME

ED 051 251 24 TE 499 828

AUTHOR	Radocy, Rudolf E.
TITLE	Development of a Test for the Nonperformance Aspects of Music Education. Final Report.
INSTITUTION	Pennsylvania State Univ., University Park.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C. Bureau of Research.
REPORT NO	R-38
BUREAU NO	BR-0-B-004
PUB DATE	Feb 71
GRANT	OEG-2-700018(509)
NOTE	149p.
EDRS PRICE	EDRS Price MF-$0.65 HC-$6.58
DESCRIPTORS	College Students, Comparative Analysis, *Computer Programs, *Criterion Referenced Tests, *Music Education, *Performance Tests, State Universities

ABSTRACT

The development of a prototype computerized, criterion-referenced test of certain nonperformance musical behaviors for administration to entering students in music education at a state university was undertaken. After the formulation of statements of competencies tc serve as conceptual bases for the formulation of test items, items were constructed for 12 subtests. Four subtests were selected for programming. The test was programmed sequentially for the IBM 1500 Instructional System. The students' score for each subtest was the number of items actually answered correctly plus the number of items for which a correct answer was assumed. The computerized test was administered to 32 undergraduate music education students. A parallel conventional version of the test was given to 28 other students, and a comparative analysis was made. The tests were not shown to be equivalent. Quantitative inadequacies may be explained by the discrepancies between estimated orders of item difficulty and the true orders of item difficulty for the particular students tested. From a qualitative standpoint, the computerized test performs adequately. With refinement, it could provide a convenient, rapid assessment of students in regard to certain expected nonperformance musical competencies. (Author/CK)

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

BR 0-B-004
PA 24

TE10

ED051251

FINAL REPORT

Project No. 0B004

Grant No. OEG-2-700018(509)

DEVELOPMENT OF A TEST FOR THE NONPERFORMANCE ASPECTS OF MUSIC EDUCATION

February 1971

U. S. Department of
Health, Education, and Welfare

Office of Education

Bureau of Research

R-38

TE 499 828

ED051251

Final Report

Project No. 0B004

Grant No. OEG-2-700018(509)

Development of a Test for the Nonperformance Aspects of Music Education

Rudolf E. Radocy

The Pennsylvania State University
University Park, Pennsylvania

February 1971

The research reported herein was performed pursuant to a grant under the Office of Education, U. S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions do not, therefore, necessarily represent official Office of Education position or policy.

U. S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education

Bureau of Research

ACKNOWLEDGMENTS

The writer wishes to express his gratitude to Dr. Frances M. Andrews for her sponsorship of the proposal and her assistance in choosing areas for testing, obtaining subjects, and preparing reports. Although she had extensive responsibilities both as Head of the Department of Music Education at The Pennsylvania State University and President of the Music Educators National Conference, she was always available to consult with the researcher. Dr. Ned C. Deihl is also acknowledged for sharing his experience with computer-assisted instruction in music education. The other members of the writer's doctoral committee, Dr. Helen I. Snyder, Dr. William Rabinowitz, Dr. James W. Dunlop, and Dr. J. David Boyle are acknowledged for their assistance during various phases of the project. Mr. Terry A. Bahn, systems operator in the Penn State CAI Laboratory, and Mr. Karl G. Borman, systems manager, gave all necessary technical assistance. Appreciation is also extended to Dr. Harold E. Mitzel, Dr. Keith A. Hall, and Mrs. Betta Kriner for administrative assistance, and to Mrs. Kris Sefchick for her considerable secretarial help.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS ii

LIST OF TABLES. vi

SUMMARY viii

Chapter

I. INTRODUCTION. 1

PURPOSE OF THE STUDY. 1

BACKGROUND. 1

INADEQUACY OF PRESENT TESTING TECHNIQUES. 3

NEED FOR CRITERION-REFERENCED TESTING 4

NEED FOR THE APPLICATION OF COMPUTER TECHNOLOGY 6

SUMMARY OF THE BACKGROUND OF THE STUDY. 7

II. REVIEW OF SELECTED LITERATURE 9

CERTAIN PRIOR DEVELOPMENTS REGARDING TESTING IN NONPERFORMANCE AREAS 10

CRITERION-REFERENCED TESTING. 14

COMPUTER TECHNOLOGY AND TESTING 19

SUMMARY OF SELECTED LITERATURE. 22

III. MATERIALS AND PROCEDURES. 24

DEVELOPMENT AND FORMULATION OF OBJECTIVES 24

Importance of Objectives. 24

Selection of Objectives 26

DEVELOPMENT OF ITEMS. 33

Selection of Musical Materials. 33

Item Construction 33

PREPARATION FOR EMPIRICAL TRIALS AND PROGRAMMING. 39

Chapter Page

EMPIRICAL TRIALS OF TEST ITEMS 40
    Necessity to Establish Item Difficulty Indices . . . 40
    Preparation of Paper-and-Pencil Forms. 41
    Administration of Items. 41
    Postadministration Analysis. 48
PROGRAMMING THE COMPUTERIZED TEST. 51
    IBM 1500 Instructional System. 51
    Programming Strategy 53
    Scoring Procedure. 54
    Audio Preparation. 55
    Film Preparation 55
    Debugging. 56
FINAL ADMINISTRATION 57
    Student Population 57
    Administrative Procedure, Computerized Version . . . 58
    Administrative Procedure, Conventional Version . . . 59
    Plan for Analysis of Data. 59
SUMMARY STATEMENT OF METHOD OF TEST DEVELOPMENT. . . . 60
IV. RESULTS AND FINDINGS 61
PRELIMINARY DATA 61
    Computation of Item Difficulty Indices 61
    Actual-Hypothetical Comparisons. 62
    Administration at Varying Institutions 69
DATA FROM FINAL ADMINISTRATION 71
    Medley Procedure 71
    Comparison of Item Difficulty Rankings 77

Chapter | Page

Comparison of Test Performance of Upper-term and Lower-term Students 81

Questionnaire Results 83

NONQUANTITATIVE FINDINGS. 92

V. SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS 96

SUMMARY 96

Objectives. 96

Test Items. 97

Programming 98

Administration and Findings 99

CONCLUSIONS 101

RECOMMENDATIONS 101

BIBLIOGRAPHY 104

APPENDICES

A. SAMPLE ITEMS. 109

B. SUMMARY TABLES FOR MEDLEY PROCEDURE DATA. 115

C. QUESTIONNAIRE ITEMS APPENDED TO BOTH TEST VERSIONS. . . . 128

D. EXAMPLE OF COURSEWRITER PROGRAMMING 131

E. EXAMPLE OF STUDENT RECORDS. 135

F. SELECTION OF TESTS FOR PROGRAMMING. 137

LIST OF TABLES

Table Page

1. Item Difficulty Indices of Selected Items for Twenty-item Scales 63

2. Descriptive Statistics Based upon Comparisons of Hypothetical and Actual Answer Strings 66

3. Rank-order Coefficients of Correlation and Number of Students Tested per Institution for Tests Administered at More Than One Institution. 70

4. Medley Procedure for ON Test, All Students 72

5. Summary of Medley Procedure Applications, Indicating Presence of Equivalence of Tests According to Four Criteria. 75

6. Discrepancies Between Estimated and Observed Item Difficulties 79

7. Rank-order Coefficients of Correlation for Difficulty Rankings. 80

8. Comparison of Upper-term and Lower-term Mean Scores. . . . 82

9. Questionnaire Responses Regarding Most Difficult Section. 84

10. Questionnaire Responses Regarding Least Difficult Section. 84

11. Questionnaire Responses Regarding Quality of Sound Reproduction 87

12. Questionnaire Responses Regarding Quality of Notation. . . 88

13. Questionnaire Responses Regarding Speededness of Test. . . 90

14. Questionnaire Responses Regarding Perceived Pressure and Tension 91

15. Questionnaire Responses Regarding Preference of Testing Situation. 93

16. Medley Procedure for Omitted Notes, All Students 116

17. Medley Procedure for Omitted Notes, Lower-term Students Only. 117

Table Page

18. Medley Procedure for Omitted Notes, Upper-term Students Only. 118

19. Medley Procedure for Overall Rhythmic Inaccuracies, All Students 119

20. Medley Procedure for Overall Rhythmic Inaccuracies, Lower-term Students Only 120

21. Medley Procedure for Overall Rhythmic Inaccuracies, Upper-term Students Only 121

22. Medley Procedure for Faulty Interpretations, All Students 122

23. Medley Procedure for Faulty Interpretations, Lower-term Students Only 123

24. Medley Procedure for Faulty Interpretations, Upper-term Students Only 124

25. Medley Procedure for Historical Classification, All Students 125

26. Medley Procedure for Historical Classification, Lower-term Students Only 126

27. Medley Procedure for Historical Classification, Upper-term Students Only 127

SUMMARY

Purpose

The purpose of the study was to develop a prototype computerized, criterion-referenced test of certain nonperformance musical behaviors for administration to entering students in music education at The Pennsylvania State University, with the expectation that the test could provide a pattern for development in similar situations.

Procedures

After the formulation of statements of competencies to serve as conceptual bases for the formulation of criterion-referenced test items, test items were constructed for twelve subtests. A total of 783 such items were constructed for twelve subtests. A total of 783 such items were administered to music and music education undergraduates at seven Pennsylvania institutions of higher education, including Penn State. Item difficulty indices were computed, and twenty-item scales, arranged in order of difficulty, were selected for each subtest. Four subtests were selected for programming.

The test was programmed for the IBM 1500 Instructional System in a sequential or incremental manner. In accordance with the programming strategy adopted for the final administration, a student began a subtest with the fourth item of the twenty-item scale. A correct response branched the student ahead to the eighth item, the assumption being that the student would have answered the first, second, and third items correctly because they were of less difficulty than the fourth item. The student proceeded in increments of four until the twentieth item was answered correctly or an initial erroneous response occurred. An initial error caused a reverse branch of three items. From that point, the

student continued the subtest in linear fashion until the end of the subtest was reached, three erroneous responses occurred in succession, or a total of five erroneous responses had occurred. The student's score for each subtest was the number of items actually answered correctly plus the number of items for which a correct answer was assumed.

Results and Conclusions

In October, 1970, the computerized test was administered to thirty-two undergraduate music education students at Penn State. A parallel conventional version of the test was administered to twenty-eight other students, and the two versions were compared with an analysis-of-variance procedure for equivalency. The tests were not shown to be equivalent, although their mean scores did not, with one exception, differ significantly. Quantitative inadequacies may be explained by the discrepancies between estimated orders of item difficulty and the true orders of item difficulty for the particular students tested.

From a qualitative standpoint, the computerized test performs adequately. Refinement is indicated by reordering of the test items on the basis of estimates of item difficulty obtained from larger groups of students. Lengthening the test to include areas representative of more behaviors might also be in order. With such refinement, the test could provide a convenient, rapid assessment of the status of music education students in regard to certain expected nonperformance musical competencies.

CHAPTER I

INTRODUCTION

PURPOSE OF THE STUDY

The basic purpose of this study was to develop a prototype computerized criterion-referenced test for measuring competencies in certain nonperformance musical behaviors present in undergraduate students commencing their course of study in music education. The prototype was developed utilizing students and resources of The Pennsylvania State University at University Park, Pennsylvania, and six other Pennsylvania institutions of higher education.

BACKGROUND

College students pursuing a course of study in music or music education include in their program the study of nonperformance areas, i.e., areas such as music theory, music history, and music literature which are not directly concerned with vocal or instrumental performance. Adequate musical preparation for entry into the profession of music education involves more than the development of technical vocal and instrumental skills. The formal music education necessary for the prospective teacher and performer should include thorough theoretical, historical, and stylistic study.[1]

The standards and expectancies of colleges and universities regarding competence in nonperformance areas vary; learners vary. If a particular college music or music education department could reliably measure its own entering students' nonperformance musical behaviors and compare the measurements with the particular expectancies of the college, certain curricular problems might be alleviated. Needed remedial learning experiences for those students identified as not meeting minimal expectancies could be indicated. Qualitative descriptions and analyses of nonperformance musical behaviors could be a basis for advanced course placement and exemption from certain courses.

---

[1] James Jorgenson, "Advice to the Potential College Music Major," Instrumentalist, XXII (April, 1968), 38-39.

Although nothing in this area had been done prior to the research reported herein, it appeared that a computer-based instructional system, designed for rapid processing of student responses to interrogative stimuli, could serve as a means of measuring with speed, flexibility, and efficiency the extent to which expectancies in nonperformance musical behaviors were met by a given student. Description and analysis of student nonperformance musical behaviors could be facilitated by programming a computer to serve as a device for the measurement of proficiency in such behaviors.

Given sufficient breadth and depth of observation, such a measuring device could serve as a diagnostic achievement test because it would purport to measure a certain pattern of musical achievement. The current lack of music tests which serve as diagnostic tools has been cited by Lehman.[2] At present, it is unlikely that existing published music tests adequately serve as a diagnostic achievement test for comparison of observed nonperformance musical behaviors with expected nonperformance musical behaviors.

---

[2] Paul R. Lehman, Tests and Measurements in Music (Englewood Cliffs, New Jersey: Prentice Hall, Inc., 1968), p. 86.

INADEQUACY OF PRESENT TESTING TECHNIQUES

Test items intended to measure musical behavior of students at a particular institution of higher learning should be based upon the goals, standards, and criteria for success in that institution. The particular objectives of one institution may be considered to be somewhat different from another. Tests for similar purposes in various music and music education departments may have similar formats, but content and sequencing of items should be free to vary. It is apparent that tests prepared on a national or regional basis with a rigid content and order of items may fail to reflect the instructional objectives and emphases of a particular music or music education faculty.

Music educators lack a national consensus as to what specific musical outcomes are expected as a result of instruction in music. No existing published achievement test is likely to receive widespread acceptance as a measurement tool because the profession does not appear to have a sufficient degree of consensus with regard to what musical behaviors are desirable.[3] Consequently, it was proposed to begin the development of the proposed measuring device by constructing a test of certain nonperformance musical behaviors for a specific institution where a consensus of institutional goals was obtainable. The pattern of development that has evolved is adaptable for application elsewhere.

---

[3] Lehman, pp. 57-58.

NEED FOR CRITERION-REFERENCED TESTING

It was proposed to develop a test that would assess nonperformance musical behaviors in relation to criterion behaviors. The behaviors to be observed and measured were to be specified and stated in the form of observable student objectives. The original intent was that these objectives were to represent the minimal amount of competence that entering music or music education students at a particular institution could be expected to display as evidence of criterion attainment. Although the objectives were eventually expressed in terms of observable competencies which an undergraduate student in music education should attain in the course of his pre-professional training, rather than in numerical expressions of desirable entering competencies, the specification of the behaviors to be measured as the initial phase of test development was in accordance with contemporary principles of test development.[4]

The distinction between norm-referenced and criterion-referenced measures is vital and of fundamental importance. Glaser explains that two kinds of primary information, differing principally in the standard used as a reference, are obtainable from an achievement test. The relative ordering of individuals with respect to their test performance is the type of information provided by a norm-referenced measure, a measure dependent upon a relative standard for relating individuals to each other. Another type of information is the degree to which the student has attained criterion performance and is provided by a criterion-referenced measure which is dependent upon an absolute standard of quality to represent criterion performance.[5]

---

[4] Robert Glaser, "Instructional Technology and the Measurement of Learning Outcomes: Some Questions," American Psychologist, XVIII (August, 1963), 520; Robert Glaser and David J. Klaus, "Proficiency Measurement: Assessing Human Performance," Psychological Principles in System Development, Robert M. Gagné, editor (New York: Holt, Rinehart, and Winston, Inc., 1962), p. 430; C. M. Lindvall, Testing and Evaluation: An Introduction (New York: Harcourt, Brace, and World, Inc., 1961), pp. 23-25.

A criterion-referenced measure provides explicit information regarding an individual's ability to perform a task. The individual's score indicates the degree of competence he has attained in relation to an ordered continuum of expected behaviors rather than in relation to the performance of others.[6]

A norm-referenced test compares individuals with each other rather than with a behavioral standard; it indicates only how much a student knows with respect to other students. The shortcomings of ordinary norm-referenced achievement tests for assessment of learning have been recognized by various authorities in regard to the ongoing national assessment program.[7]

Although Cronbach defines a test as "a systematic procedure for comparing the behavior of two or more persons,"[8] the comparison of one person to another was not the purpose of the test developed in this study. Rather, the purpose was perceived as a comparison of a person's existing observed nonperformance musical behaviors with desired nonperformance musical behaviors as represented by test items that demonstrate attainment of criterion competencies, i.e., a criterion-referenced test.

---

[5] Glaser, 519.

[6] Glaser and Klaus, p. 422.

[7] Caroline Hightower, How Much Are Students Learning? Plans for a National Assessment of Education (Ann Arbor, Michigan: The Committee on Assessing the Progress of Education, 1968), p. 6.

[8] Lee J. Cronbach, Essentials of Psychological Testing (2nd ed.; New York: Harper and Row, 1960), p. 21.

Tests which presently exist in music, although meeting Cronbach's definition, do not appear to compare observed with expected behavior. This is not unexpected since the assessment procedures conventionally used in development of the typical standardized test in any area of knowledge do not include a method for assessing student performance in terms of instructional objectives. Existing achievement tests appear to have as their purpose the demonstration of the great range of individual differences in behavior. Continued refinement of norm-referenced tests to maximize their discriminatory power is not likely to be worthwhile for the purpose of measuring achievement in terms of expected behaviors.[9] Comparison of the observed with the expected requires criteria for that which is expected, not discrimination among those who are observed.

NEED FOR THE APPLICATION OF COMPUTER TECHNOLOGY

Technological developments have made it possible to create new testing materials and present them in a variety of ways. A computer may be programmed to present varied test stimuli, to record and evaluate the responses, and to provide a printed summary and interpretation of each individual test performance in relation to a behavioral standard. Tyler states:

---

[9] Robert M. Gagne', The Conditions of Learning (New York: Holt, Rinehart, and Winston, Inc., 1966), p. 258.

> Now that high-speed computers and electronic data processing make individual diagnosis, recording, and treatment feasible, teachers do not have appropriate evaluation instruments to guide greater individualization of instruction. We are still so obsessed with the ranking of individuals on the basis of scores that we have not developed adequately the tools and procedures required. Theory and practice need to be reexamined in terms of present conditions and opportunities.[10]

It was the researcher's belief that computer technology could be used effectively to bring new techniques to bear upon the problem of the measurement of nonperformance musical behaviors. The technique of sequential or incremental testing, whereby the student's response history is utilized to determine the order of presentation of test items to an individual student, appeared particularly promising. Furthermore, the computer can smoothly and rapidly present a variety of musical stimuli in an individualized manner by coordinating the appropriate auxiliary apparatus.

SUMMARY OF THE BACKGROUND OF THE STUDY

A lack of a suitable measuring instrument was perceived for comparing certain nonperformance musical behaviors of entering college music and music education students with expected levels of competence. Norm-referenced tests that discriminate between individuals were viewed to be inappropriate for the purpose. It was therefore proposed to use computerized presentation and analysis to rapidly administer a criterion-referenced test to evaluate the behavior of entering freshman music and music education majors in certain nonperformance areas in relation to defined expectancies.

---

[10] Ralph W. Tyler, "Changing Concepts of Educational Evaluation," Perspectives of Curriculum Evaluation, Ralph W. Tyler, Robert M. Gagne, and Michael Scriven, editors (Chicago: Rand McNally and Company, 1967), p. 17.

CHAPTER II

REVIEW OF SELECTED LITERATURE

The purpose of this chapter is to provide a conceptual basis for the work that was undertaken by illustrating research and opinion that had been previously applied to the development of tests in nonperformance areas for entering students, criterion-referenced measures, and feasibility of computerized testing.

A substantial amount of literature has been developed regarding tests and measurements in music. Lehman[1] and Whybrew[2] have written textbooks discussing problems inherent in music testing, certain statistical concepts, the classification of tests as aptitude or achievement measures, and published standardized tests in music. As portions of psychology of music texts, psychologists such as Farnsworth[3] and Lundin[4] have reviewed tests and discussed problems in the context of definition and measurement of musical behavior. The controversy between the Seashore atomistic view of musical talent and the Mursell general view of musical talent with implications for testing has been widely reported.[5] A comprehensive listing of literature pertinent to the general topic of testing in music, compiled and categorized by Lehman, provides a total of 298 discrete entries, including psychological tests, reviews, texts containing sections on music tests, experimental studies, studies of published tests, and works regarding the status of testing.[6]

---

[1] Paul R. Lehman, Tests and Measurements in Music (Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1968), pp. 1-99.

[2] William K. Whybrew, Measurement and Evaluation in Music (Dubuque, Iowa: The William C. Brown Company, 1962), pp. 1-184.

[3] Paul R. Farnsworth, The Social Psychology of Music (New York: Holt, Rinehart, and Winston, Inc., 1958), pp. 1-304.

[4] Robert W. Lundin, An Objective Psychology of Music (2nd ed.; New York: Ronald Press, 1967), pp. 1-345.

[5] Lehman, pp. 40-41.

General tests and measurements literature, though related, is not as directly pertinent to the developmental research reported herein as are certain more specific materials. Literature regarding standardized tests of music, musical aptitude and its definition, and the philosophical justification for testing is only peripherally related to the conceptual basis of the research. The literature critical to the present study has been devoted to (1) development of tests for diagnosis of difficulties of entering music and music education students in nonperformance areas, (2) the feasibility of the proposed computerized approach, and (3) criterion-referenced measures.

CERTAIN PRIOR DEVELOPMENTS REGARDING TESTING IN NONPERFORMANCE AREAS

Ball developed a test measuring responses to elements of rhythm, melody, and harmony, singly and in combination, to serve as a college entrance test of music. The items were administered for trial purposes to equal samples of high musical ability and low musical ability students, with theory grades and teachers' ratings as the basis of ability determination. The final test items were selected on the basis of their power to discriminate between the high and low groups, rather than on the basis of how well they represented expected criterion performance. Ball's test does not appear to be criterion-referenced when his procedure for item selection is considered.[7]

---

[6] Paul R. Lehman, "A Selected Bibliography of Works on Music Testing," Journal of Research in Music Education, XVII (Winter, 1969), 428-442.

Perry constructed a test to be administered to entering freshmen for purposes of guiding, counseling, placing, and selecting the students in and for theory classes. After one semester of theory instruction, a correlation coefficient of .60 was found between scores on seven selected predictor portions of the Perry test and criterion scores obtained from proficiency examinations in theory. Perry's purpose was to make a comparison of the abilities of various predictive measures to predict examination scores rather than comparing observed behaviors with criterion performance.[8]

Mansur devised a Wind Instrumentalist Inventory Scale for use as a paper-and-pencil objective test of achievement related to musical performance. He suggested that it could be administered to entering freshmen as a predictive and screening device for college and university instrumental groups. This performance-related test discriminates between individuals rather than ascertaining the extent to which the objectives of an institution have been met.[9]

---

[7] Charles Hershel Ball, "The Application of an Empirical Method to the Construction of a College Entrance Test in Music" (unpublished doctoral dissertation, George Peabody College for Teachers, 1964), Dissertation Abstracts, XXVI (July-August, 1965), 404.

[8] William Wade Perry, "A Comparative Study of Selected Tests for Predicting Proficiency in Collegiate Music Theory" (unpublished doctoral dissertation, North Texas State University, 1965), Dissertation Abstracts, XXVI (January, 1966), 3995-3996.

[9] Paul Max Mansur, "An Objective Performance-Related Music Achievement Test" (unpublished doctor's dissertation, The University of Oklahoma, 1965).

The Gordon Musical Aptitude Profile, a norm-referenced, published, standardized test of musical aptitude,[10] was used by Hatfield to diagnose tonal and rhythmic strengths and weaknesses in a correlational study using South Dakota State University band students. The highest intercorrelations were found between the "Tonal Imagery" section of the Gordon test and certain tonal-creative behaviors related to instrumental performance; the rhythmic results were not as clear. Criterion behaviors appropriate to band students apparently were not taken into account.[11]

Edwin Gordon, the author of the Musical Aptitude Profile, maintains that the instrument can be used to help college music administrators and teachers in the diagnosis of individual musical strengths and weaknesses.[12] In the measurement of nonperformance musical behaviors with this norm-referenced measure, however, the comparison is between observed behavior and norms based upon the test performance of a representative sample of subjects. Although this may be of some value, it is not identical to using a criterion-referenced measure. Furthermore, the use of an aptitude measure such as the Gordon test, designed to predict or forecast over an extended period of time, is somewhat questionable for diagnosis of present strengths and weaknesses in relation to a current instructional process.[13]

---

[10] Edwin Gordon, Musical Aptitude Profile (Boston: Houghton Mifflin Company, 1965).

[11] Warren Gates Hatfield, "An Investigation of the Diagnostic Validity of the Musical Aptitude Profile with Respect to Instrumental Music Performance" (unpublished doctoral dissertation, The University of Iowa, 1967), Dissertation Abstracts, XXVII (January-February, 1968), 3210A.

[12] Edwin Gordon, "Implications for the Use of the Musical Aptitude Profile with College and University Freshman Music Students," Journal of Research in Music Education, XV (Spring, 1967), 34.

Douglas grouped freshman music majors at the University of Georgia in the fall of 1964 into a tripartition of high, median, and low. The high group immediately began the study of music theory, while the median and low groups received one and two quarters of preparatory instruction respectively. Douglas found that a greater percentage of students could ultimately cope with theory as a result of being grouped, and suggested that the combination of tests be used to make the tripartition, his own test plus the Aliferis Music Achievement Test (College Entrance Level), could be useful for counseling purposes.[14]

The Aliferis test consists of six subtests: "Melodic Elements," "Melodic Idioms," "Harmonic Elements," "Harmonic Idioms," "Rhythmic Elements," and "Rhythmic Idioms." All items require some form of aural-visual discrimination, i.e., the student relates what he hears to an array of visual stimuli. Such discriminatory skills are helpful in the study of music theory; the Aliferis test was undoubtedly useful in making Douglas's tripartition. But it is a norm-referenced standardized test. The manual carefully presents norms for each section of the test, regional norms, national norms, and norms for various types of institutions.[15] Hence, the students in the Douglas study were compared with each other. A criterion-referenced measure could have been used to compare the students with University of Georgia theory standards, assuming that the standards could have been stated in a manner conducive to the construction of test items.

---

[13] Robert Glaser, Evaluation of Instruction and Changing Educational Models, C.S.E.I.P., Occasional Report No. 13 (Los Angeles: University of California at Los Angeles, Center for the Study of Evaluation of Instructional Programs, 1968), pp. 12-13.

[14] Charles Herbert Douglas, "Measuring and Equalizing Music Theory Competence of Freshmen College Music Majors" (unpublished doctoral dissertation, The Florida State University, 1965), Dissertation Abstracts, XXVI (February, 1966), 4712.

The tests developed by Ball, Perry, and Douglas are representative of the usual measuring instrument constructed for the purpose of measuring musical behaviors of entering students. Test items are selected on the basis of their powers of discrimination rather than on the basis of their relationship to pertinent criteria of performance. No criterion-referenced measure for the purpose of measuring nonperformance musical behaviors was known to the researcher at the onset of the test development reported herein. There was, however, significant interest in criterion-referenced testing outside of the field of music.

CRITERION-REFERENCED TESTING

The distinction between criterion-referenced and norm-referenced tests is made by Glaser in terms of differing kinds of primary information obtainable from the two forms of tests. Criterion-referenced measures provide information regarding the degree to which criterion performance has been attained; norm-referenced measures provide information regarding the relative ordering of individuals in terms of their observed achievement.[16] Popham and Husek clarify the distinction by explaining that norm-referenced measures generally imply a concern for selectivity, while criterion-referenced measures imply a concern for competence in an individual or the efficacy of a treatment.[17]

---

[15] James Aliferis, Aliferis Music Achievement Test (College Entrance Level) (Minneapolis, Minnesota: University of Minnesota Press, 1954).

Glaser and Klaus discuss criterion-referenced measures in relation to job training. They refer to a continuum of skill at a given task that ranges from no proficiency at all to perfect performance. The behaviors which an individual displays during testing of this skill fall at some point on the skill continuum, and the degree to which these behaviors resemble desired or criterion behaviors can be assessed by a criterion-referenced measure. Criterion levels are also ordered on a continuum; they can be established at any point where it is necessary to obtain information as to the adequacy of an individual's learning. Specific behaviors expected at a given level of proficiency, such as the college entrance level, may be identified and used to describe specific tasks which the individual is to perform.[18]

---

[16] Robert Glaser, "Instructional Technology and the Measurement of Learning Outcomes: Some Questions," American Psychologist, XVII (August, 1963), 520.

[17] W. James Popham and T. R. Husek, "Implications of Criterion-Referenced Measurement," Journal of Educational Measurement, VI (Spring, 1969), 1-9.

[18] Robert Glaser and David J. Klaus, "Proficiency Measurement: Assessing Human Performance," Psychological Principles in System Development, Robert M. Gagné, editor (New York: Holt, Rinehart, and Winston, Inc., 1962), pp. 421-422.

The lack of consensus among music educators as to what musical outcomes are to be expected as a result of instruction does not mean that criteria for a criterion-referenced test cannot be selected. Arbitrary standards may be established by the faculty of a given institution with regard to their own philosophy, experience, and view of music education. Glaser and Klaus state:

> . . .the lack of well-defined system standards does not preclude the use of criterion-referenced measures. Arbitrary proficiency levels can be established for minimum performance. For instance, it is possible to select standards in academic training which reflect decisions as to the least amount of end-of-course comptenecy the student is expected to attain . . .it is possible to use the maximum amount of course content presented to the student as a standard.[19]

A conceptual basis for criteria and objectives of a criterion-referenced test in music may be found in an Interim Report of the MENC Commission on Teacher Education, wherein the Commission states certain competencies that should be displayed by qualified music educators as a result of their teacher training experience. The Commission indicates that all music educators should display skills in performance, composition, and analysis. Of particular importance for the research reported herein is the Commission's endorsement of competency in the supervision and evaluation of the performance of others and competency in the identification of compositional devices. The researcher's test of certain nonperformance musical behaviors was completed and administered prior to the appearance of the report; however, future work may draw increasingly upon the Commission's publication.[20]

---

[19] Glaser and Klaus, p. 426.

Once criteria and objectives related to skills or competencies are established, it may be desirable to obtain information about an individual's degree of skill or competency. Norm-referenced measures do not provide much information regarding individual degrees of skill or competency; they provide comparisons between a particular individual's test performance and the performance of other members of his group.[21]

Norm-referenced tests suggest grouping those who are tested into a normal distribution. Bloom notes that although the normal distribution is the distribution most appropriate to chance and random activity, education is a purposeful activity. The distribution of student achievement, therefore, should be quite different from the normal distribution if teachers are effective in their instruction. Relative standards are inappropriate if teachers desire to bring all their students to a criterion level.[22]

Glaser indicates that criterion-referenced tests do not group students into a normal distribution. Such tests provide individual information independent of reference to the performance of others because criterion-referenced tests indicate the correspondence between an individual's observed behavior and an underlying continuum of achievement.[23]

---

[20] MENC Commission on Teacher Education, "Teacher Education in Music: An Interim Report of the MENC Commission on Teacher Education," Music Educators Journal, LVII (October, 1970), 38-41.

[21] Glaser and Klaus, p. 422.

[22] Benjamin S. Bloom, "Learning for Mastery," Evaluation Comment, I (May, 1968), 2-3.

Popham and Husek discuss differences between criterion-referenced measures and norm-referenced measures in terms of item selection; they state that the writer of norm-referenced measures, in an effort to promote variant scores for the purpose of discriminating among individuals, rejects test items that are quite difficult or quite easy. The writer of the criterion-referenced measure is concerned with whether or not the test items represent the desired class of behaviors.[24] The inappropriateness of deliberately promoting a spread of scores when one is concerned with group achievement of criterion behaviors is also discussed by Glaser and Cox,[25] while Cox and Vargas suggest that item selection for a criterion-referenced measure may be more profitably conducted by evaluating items through a pretest-posttest method to determine the items' ability to indicate whether or not instruction benefited the student.[26] An item with a difficulty index of 0.00 or 1.00 might be rejected as nondiscriminating by the writer of the norm-referenced test, but such an item on a criterion-referenced test may be clear evidence that a criterion behavior has or has not been attained.

---

[23] Glaser, "Instructional Technology and the Measurement of Learning Outcomes: Some Questions," 519-520.

[24] Popham and Husek, 4.

[25] Robert Glaser and Richard C. Cox, "Criterion-Referenced Testing for the Measurement of Educational Outcomes," Instructional Process and Media Innovation, Robert A. Weisgerber, editor (Chicago: Rand McNally and Co., Inc., 1968), p. 549.

[26] Richard C. Cox and Julie S. Vargas, "A Comparison of Item Selection Techniques for Norm-referenced and Criterion-referenced Tests" (paper read at the annual meeting of the National Council on Measurement in Education, February, 1966, Chicago).

COMPUTER TECHNOLOGY AND TESTING

The growth and increasing sophistication of computer technology in recent years has major applications to testing. Entire computer-based test development systems are feasible, both in schools and industry. Tests of the criterion-referenced and norm-referenced variety can be developed, presented, and analyzed at very rapid speeds.[27]

The computerized presentation and analysis of a test initially constructed off-line (i.e., without a computer) is perhaps less sophisticated than computerized construction of a test from a vast bank of potential items, but such presentations have been successfully developed. Greer, for example, conducted a pioneering study of the use of a computer to score and analyze a test and prepare a diagnostic report. He concluded that computerized testing was feasible, and that it increased efficiency and provided useful basic information at the United States Naval Examining Center. It was recommended that educators consider the computer for scoring, analysis, and diagnosis.[28] Williams found the computer to be valuable for individual diagnosis and evaluation in a reading program.[29]

---

[27] Jack V. Edling, "New Media Applications," Man-Machine Systems in Education, John W. Loughary, editor (New York: Harper and Row, 1966), p. 76.

French developed a means of rapidly presenting and scoring test items, sequentially arranged according to difficulty, for vocational and technical students through the IBM 1050 computerized typewriter terminal and 1410 computer. Numerical and verbal items were selected from the Henmon-Nelson Tests of Mental Ability. Rather than presenting every item to every student, French utilized an individualized branching approach. The numerical test items were presented in order of increasing difficulty in increments of eight; i.e., a student was asked to respond to every eighth item. An incorrect response caused the student to go back five times in the test program and be presented with every second item. A second incorrect response branched the student back five items and presented every item, omitting items that were previously presented. Four misses out of seven items discontinued the test program.[30]

---

[28] Harry Holt Greer, Jr., "The Application of a Digital Computer to Scoring and Analysis of Examinations and the Preparation of Diagnostic Reports" (unpublished doctoral dissertation, The George Washington University, 1966), Dissertation Abstracts, XXVII (September-October, 1966), 923A.

[29] Gilbert Williams, "The Use of the Computer for Testing, Programming, and Instruction," Research in Education, III (May, 1958), 195.

[30] Joseph L. French, "Numerical and Verbal Aptitude Tests Administered at the CAI Student Station," Semi-Annual Progress Report (prepared by Harold E. Mitzel, et al), Experimentation with Computer-Assisted Instruction in Technical Education, Project No. 5-85-074. (University Park, Pa.: The Pennsylvania State University Computer-Assisted Instruction Laboratory, 1967), pp. 59-62.

The items were arranged in a linear order of difficulty in French's test. The student commenced the test with an easy item and gradually worked toward the difficult items. An alternative arrangement was utilized by Hansen, who programmed test items from a midterm physics examination at Florida State University for a computerized presentation. The student commenced the Hansen test by responding to an item in the middle of the difficulty scale. A correct response branched the student to a harder item; an incorrect response branched the student to an easier item. The student always moved ahead, but the difficulty of the next item presented was determined by his response to the present item.[31]

A concept of sequential testing is illustrated by the French and the Hansen tests. In each case, the test items are arranged in a purposeful nonrandom sequence. The use of the computer made it possible for a student to substantially complete each test by taking only certain items, depending upon his response history. A computer is not essential to a sequential test if every student is to respond to every item; Cox and Graham developed a sequential test based on a sequence of arithmetic behaviors ordered according to a hierarchy of difficulty upon which the ability to add two two-digit numerals involving "carrying" appeared to be based.[32]

---

[31] Duncan N. Hansen, An Investigation of Computer-Based Science Testing, FSU CAI Center, Semiannual Progress Report, Report No. 6 (prepared by Duncan N. Hansen, Walter Dick, and Henry T. Lippert) (Tallahassee, Florida: Florida State University Computer-Assisted Instruction Center, 1968), pp. 59-94.

[32] Richard C. Cox and Glenn T. Graham, "The Development of a Sequentially Scaled Achievement Test," Journal of Educational Measurement, III (Summer, 1966), 147-150.

In connection with other research with computer-assisted instruction, tests have been utilized as part of the instructional process to determine what sections of a computer-assisted course might be of most benefit to the student. For example, Lippert and Ehlers developed for computerization a set of test items reflecting competencies which an entering graduate student in the social science area was believed to require. These items were used to plot computer-assisted instruction for the areas of weakness revealed by the test.[33] Deihl programmed a diagnostic quiz at the beginning of the rhythm section of a computer-assisted instruction course in certain skills of instrumental music, developed with the assistance of the researcher. Based upon the student's quiz performance, a decision was made to branch the student through one or two remedial sections or to branch him directly to the rhythm program.[34]

SUMMARY OF SELECTED LITERATURE

Examination of pertinent literature indicates that tests developed in recent years to measure entering musical behaviors in nonperformance areas tend to be useful principally for the separation of entering students into groups. Criterion-referenced testing has not been investigated in the area of music. Computer technology may be utilized for rapid test administration and analysis; it is particularly useful for utilizing a student's response history in determining which test items from a sequential test are to be administered. Thus, a conceptual framework for the present research has been established.

---

[33] Henry T. Lippert and Walter Ehlers, Computer-Based Testing, FSU CAI Center, Annual Progress Report, Report No. 7 (prepared by Duncan N. Hensen, Walter Dick, and Henry T. Lippert) (Tallahassee, Florida: Florida State University Computer-Assisted Instruction Center, 1968), pp. 18-20.

[34] Ned C. Deihl, Development and Evaluation of Computer-Assisted Instruction in Instrumental Music, Project No. 7-0760, ERIC No. ED 035 314. (Washington: Office of Education, U. S. Department of Health, Education, and Welfare, 1969), p. 22.

CHAPTER III

MATERIALS AND PROCEDURES

Procedures followed in the development of the test and materials are discussed in this chapter. The stages of development included the development and formulation of objectives, development of test items, empirical trial of test items, programming, and main test administration.

DEVELOPMENT AND FORMULATION OF OBJECTIVES

Importance of Objectives

The construction of any test is impossible without some conceptualization of what is to be measured. Tests are written because test authors are seeking to determine whether or not certain expected behaviors occur. Consequently, those behaviors and the means for their recognition must be specified. In the case of achievement tests, such behaviors must be related to instruction. Glaser states that it is mandatory to specify minimum levels of achievement which indicate the minimum level of competence a student should display at any crucial point in an instructional sequence.[1] Glaser and Klaus maintain that the specification of behavior which is to be observed and measured is the initial step in the development of a measure of proficiency.[2] Lindvall stresses that any plan to assess achievement must begin with a clear specification of objectives.[3] Regarding what he perceives to be a beginning revolution in education, particularly in regard to individualization of instruction and concentration upon mastery of learning rather than discrimination among learners, Mitzel indicates that achievement tests need to be keyed to course objectives, stated in behavioral terms.[4] Kibler, Barker, and Miles believe that test preparation is simplified when evaluative measures are designed to measure the success of instruction in terms of behaviors identical to those specified in objectives.[5] Lehman maintains that the most important part of test construction is clearly defining the objectives of the test.[6] The test that was developed is criterion-referenced; Leonhard and House state, ". . .the only criteria applicable to the music program are the objectives."[7]

---

[1] Robert Glaser, "Instructional Technology and the Measurement of Learning Outcomes: Some Questions," American Psychologist, XVII (August, 1963), 520.

---

[2] Robert Glaser and David J. Klaus, "Proficiency Measurement: Assessing Human Performance," Psychological Principles in System Development, Robert M. Gagne, editor (New York: Holt, Rinehart, and Winston, Inc., 1962), p. 430.

[3] C. M. Lindvall, Measuring Pupil Achievement and Aptitude (New York: Harcourt, Brace, and World, Inc., 1967), p. 12.

[4] Harold E. Mitzel, "The IMPENDING Instruction Revolution," Phi Delta Kappan, LI (April, 1970), 438.

[5] Robert J. Kibler, Larry L. Barker, and David T. Miles, Behavioral Objectives and Instruction (Boston: Allyn and Bacon, Inc., 1970), p. 13.

[6] Paul R Lehman, Tests and Measurements in Music (Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1968), p. 79.

[7] Charles Leonhard and Robert W. House, Foundations and Principles of Music Education (New York: McGraw-Hill Book Company, Inc., 1959), p. 146.

The formulation of objectives related to instruction appears to be the necessary first step in test construction; prominent writers call attention to the importance of objective construction, and appropriate objectives seem valuable as tools for the conceptualization of what is to be measured as well as statements of criteria for the development of a criterion-referenced test. ("Instruction" here is used to represent the sum of musical input received by the student prior to the commencement of test administration, and is not limited to a particular amount of input from any formalized course situation.)

Selection of Objectives

Preparation of behavioral objectives checklist. To simultaneously state valid objectives for undergraduate students in music education and delineate criteria to determine the extent of attainment of the objectives, a checklist of forty-two objectives written in the form, "Given _________, the student will be able to _________," was prepared and distributed by the researcher to faculty members of the Department of Music Education and graduate students in music education at The Pennsylvania State University during the summer term of 1969. The forty-two statements of behavioral objectives were related to the following arbitrarily selected nonperformance musical behaviors:

Aural recognition and identification of melodic intervals.

Aural recognition and identification of harmonic intervals.

Aural recognition and classification of major, minor, augmented, and diminished triads.

Insertion of missing notes into visual notational displays of aurally perceived melodies.

Insertion of missing notes into visual notational displays of aurally perceived harmonic sequences.

Recognition of harmonically correct parts to complete four-part harmonic passages.

Construction of harmonically correct parts to complete four-part harmonic passages when one part is missing.

Recognition and location of aural-visual rhythmic discrepancies.

Selection from arrays of explanations of appropriate explanations of incorrectly performed rhythmic patterns.

Recognition and location of incorrectly notated measures for given meter signatures.

Selection of the members of pairs of examples that are performed "better" when "better" refers to tapered phrase endings, dynamics, appropriateness of breathing, or appropriateness of articulation style.

Indication of the appropriateness of overall interpretation of examples and identification of inappropriateness as being due to inappropriate tempo, inappropriate articulation, excessive rubato, lack of rubato, or inappropriate dynamics.

Classification of examples as being representative of Medieval, Renaissance, Baroque, Classical, Romantic, or Modern Periods.

Selection of the members of pairs of examples containing ornamentation (trills, grace notes, mordente, grupetti) that are performed in the more appropriate style.

Three behavioral objectives, varying in the size of the array of choices available to the student and/or the number (five, ten, or twenty) of twenty examples to which the student was to respond correctly, were constructed for each behavioral area. Respondents were asked to indicate whether they believed each objective to be appropriate for freshman music education majors, seniors, both groups of students, or neither group. The original research proposal had called for separate sets of items for entering students and students near graduation; therefore, there were separate "freshman" and "senior" categories.

Analysis of the checklist. Perhaps the checklist (to which six faculty members and twenty-nine graduate students responded) would have been more useful had the respondents been asked to rate each objective as "appropriate" or "inappropriate" for "music education students." Respondents seemed to have difficulty classifying according to freshmen and seniors.

Further difficulty in analyzing the checklist was experienced when statistical tests were considered to seek any trends in the data for each objective. The χ^2 one-sample test, originally planned, was abandoned because it shows only that observed frequencies do or do not deviate significantly from expected frequencies; what the expected frequencies should be was not clear. The Kolmogorov-Smirnov one-sample test was applied to the graduate students' data for each objective by ordering the four categories of responses on a difficulty continuum running freshmen→both→seniors→neither, but the abandonment of this statistical test appeared advisable because, although significant deviations from expected cumulative frequencies were revealed, particularly when "senior" and "neither" categories were heavily checked, considerable doubt was raised about the appropriateness of ordering essentially discrete data on a continuum. Respondents may have differed widely in their interpretation of the "both" and "neither" categories; they may not have checked them in terms of difficulty. The application of a binomial test to each objective by formulation of dichotomies of "most frequent response-all other responses" was believed to show any strong trend to one category where such a trend existed, but the small size of the faculty "sample" made the test inappropriate for that group.

Faculty opinion of any proposed objective was considered to be of prime importance. It was decided to reject any objective that two or more faculty members had checked as being inappropriate for either group. Objectives thus rejected totalled eleven; all rejected objectives had asked the student to respond correctly to twenty of twenty items. None of the fourteen categories of nonperformance musical behaviors was completely rejected; i.e., in no case were all three objectives formulated for a particular area checked as appropriate for neither group.

Qualitative analysis of faculty and student feedback was more illuminating than the attempts at statistical analysis. One frequent point raised was the difficulty of judging the appropriateness of an objective without seeing and hearing the test items to be associated with the objective. Some faculty members questioned whether the traditional tasks of interval and triad recognition were really indicative of any desirable competencies for music educators.

Choice of objectives. No behavioral area was completely rejected, and, to a certain extent, final judgment of the appropriateness of an objective appeared to depend upon the resultant test items. The behavioral objectives checklist and the behavioral areas upon which the checklist objectives had been based were reviewed; the following non-quantitative objectives for music education students were stated to provide a basis for item construction:

1. The music education major should aurally recognize and identify melodic intervals.

2. The music education major should aurally recognize and identify harmonic intervals.

3. The music education major should aurally recognize and classify major, minor, augmented, and diminished triads.

4. The music education major should insert missing notes into visual notational displays of aurally perceived melodies.

5. The music education major should recognize and locate aural-visual pitch discrepancies in four-part harmonic passages.

6. The music education major should recognize and locate aural-visual rhythmic discrepancies.

7. The music education major should select from arrays of explanations appropriate explanations of incorrectly performed rhythmic patterns.

8. The music education major should recognize and locate incorrectly notated measures for given meter signatures.

9. The music education major should select the members of pairs of examples that are performed "better" when "better" refers to tapered phrase endings, dynamics, appropriateness of breathing, or appropriateness of articulation style.

10. The music education major should identify and classify inappropriateness of interpretation when the inappropriateness is due to inappropriate tempo, inappropriate articulation, excessive rubato, lack of rubato, or inappropriate dynamics.

11. The music education major should classify examples as being stylistically representative of the Baroque, Classical, Romantic, or Modern Period.

12. The music education major should classify examples as being stylistically representative of acid rock, soul, country-western, pop standard, "bubble gum", folk, folk rock, or blues.

In its Interim Report, the MENC Commission on Teacher Education presented a broad list of musical competencies, including skills in performance, composition, and analysis, which should result from a total undergraduate program in music education. The objectives stated above are all conceptually germane to one or more of the competencies suggested by the Commission. Objectives one, two, three, four, and five, for example, may be deemed relevant to the Commission's call for competency in the identification of compositional devices and the organization of sounds for personal expression. Relevancy is apparent between the Commission's declaration that music educators need to be effective in the supervision and evaluation of the musical performance of others and objectives five, six, seven, eight, nine, and ten. Numerous other relationships may be evidenced upon comparison of the objectives and the Commission's report.[8]

The list of objectives was not intended to cover comprehensively the universe of nonperformance musical behaviors; it was intended to provide a working list of expected behaviors upon which to build test items. The ambiguity which results from the lack of numerical criteria and indication of a time and place at which the behavior should occur is intentional. The test which was constructed measures, within each area tested, the degree to which, in terms of the number of items on a scale ordered in empirically established difficulty levels, a behavior is mastered. Prior to receipt of an undergraduate degree in music education, at some point in time, a music education major, in the opinion of the researcher as substantiated by members of a music education faculty, ought to display the behaviors listed. The criterion-referencing of the test derives from the construction of items in reference to expected behaviors, rather than from specific course objectives or a series of behaviors prerequisite to a criterion behavior. For research purposes, it was deemed sufficient to construct test items in relation to the list.

---

[8] MENC Commission on Teacher Education, "Teacher Education in Music: An Interim Report of the MENC Commission on Teacher Education," Music Educators Journal, LXII (October, 1970), 39-41.

DEVELOPMENT OF ITEMS

Selection of Musical Materials

Musical materials selected for item construction included melodies chosen from pedagogical and orchestral literature, chorales, and recordings of various styles of music. Although selection of material was made with its usefulness for future test items in mind, no particular musical example was selected for any particular test item.

Item Construction

Appropriate musical excerpts were examined in light of objectives. In a broad sense, all test items ask the student either to classify or to detect a discrepancy between what he sees and what he hears. There was a conscious effort to vary the difficulty of items within each section. A variety of instruments was utilized for recording; length of excerpt and apparent saliency of aural-visual discrepancies were varied. Thirteen groups of test items were constructed and prepared for empirical trial. Scales of twenty items each were planned for computerization, but, in the initial construction stage, an excess of items was developed to increase the likelihood of obtaining satisfactory twenty-item scales.

Melodic intervals. The melodic intervals group[9] consists of seventy-eight pairs of successive tones played on piano, clarinet, bassoon, baritone, tuba, flute, oboe, bass clarinet, horn, alto saxophone, cornet, or trombone. The unison, minor second, major second, minor third, major third, perfect fourth, tritone, perfect fifth, minor sixth, major sixth, minor seventh, major seventh, and perfect octave appear six times each, with the lower tone of the pair occurring once within each of the octaves C_C - C, C - c, c - c', c' - c", c" - c'", and c'" - c"".[10] In all cases the lower tone is played first. The student's task is to choose the name of the interval from an array of twelve names. No musical notation is viewed by the student.

---

[9] Hereafter referred to as the MI group.

Harmonic intervals. The harmonic intervals group[11] is similar to the MI group. The identical intervals are utilized, played simultaneously, presented in a different order, and performed with different instrumentation. Again, the stimulus is aural.

Triad classification. Major, minor, augmented, and diminished triads are presented in the triad classification group.[12] The four types of triads appear in root position, first inversion, and second inversion with the lowest of three tones occurring once within each of the octaves C - c, c - c', c' - c", and c" - c'". The forty-eight triads are played on piano or with various combinations of three wind instruments utilizing flute, clarinet, oboe, bassoon, alto saxophone, bass clarinet, cornet, horn, trombone, baritone, or tuba. The student taking the test views no notation; after hearing a triad he is asked to indicate whether the triad is major, minor, augmented, or diminished.

---

[10] This notation is in accordance with that used in Robert W. Lundin, An Objective Psychology of Music (2nd ed.; New York: Ronald Press, 1967), p. 19.

[11] Hereafter referred to as the HI group.

[12] Hereafter referred to as the TC group.

Omitted notes. The omitted notes group[13] requires the student to follow the musical notation while he listens to a performance of the musical excerpt. One note is missing from the notational display; the "omitted" note is replaced by a question mark. After he hears one of the seventy-three ON items, the student is asked to choose from an array of four notes the note which represents the pitch he heard at the location of the question mark. Examples of ON items are found in Appendix A.

Erroneous notes. Four part chorales are used in the erroneous notes group[14]; there are eighty items in the item pool. Each chorale is performed by a woodwind group, a brass group, or a pianist. One note is performed incorrectly in seventy of the EN items, and the student is asked to indicate which one of four circled notes on the notational display is incorrectly performed. Ten items ask the student to choose from the entire display. Errors vary in assumed difficulty of detection from incorrect pitches that disagree with the key signature to changed doublings within triads.

Rhythmic discrepancies. Changes from notated rhythm occur within a measure in the seventy-three items comprising the rhythmic discrepancies group.[15] The student indicates the number of the measure containing the discrepancy, if any, between his aural and visual input.

---

[13] Hereafter referred to as the ON group.

[14] Hereafter referred to as the EN group.

[15] Hereafter referred to as the RD group.

Rhythmic errors include interchanged note values, omitted rests, incorrectly performed patterns, and doubled or halved note values. Piano and a variety of wind instruments are used to perform the items.

Overall rhythmic inaccuracies. The overall rhythmic inaccuracies group[16] differs from the RD group; in an ORI item, the rhythm problem occurs over more than one measure. The tempo or a pattern may be consistently distorted. Certain items contain no inaccuracies. Conventional multiple-choice format is used for the seventy-three items; the student chooses his answer for each item from an array of four explanations of the rhythmic inaccuracy. Appendix A contains examples of ORI items.

Incorrect measure for signature. A strictly visual incorrect measure for signature group[17] asks the student to study four-measure patterns written in one-line rhythmic notation and, for eighty items, select the one measure, if any, that contains an incorrect total of counts for the given meter signature.

Better phrasing. Two versions, labelled "A" and "B", of each of seventy-three melodies are presented to the student in the better phrasing group.[18] The notation is displayed to the student; wind instruments are used for the performance. The student's task is to indicate whether the "A" or "B" version is phrased better, or to indicate that there is no substantial difference. Most items do contain a difference; one version contains an abruptly terminated note or an unnatural interruption of the musical flow caused by inhaling at an improper time.

---

[16] Hereafter referred to as the EN group.

[17] Hereafter referred to as the RD group.

[18] Hereafter referred to as the ORI group.

Faulty interpretation. In a group of seventy-three faulty interpretation[19] items, a melody is performed on a wind instrument or piano while the student follows the notation. In the manner of printed music, the visual display contains certain information about tempo, dynamics, and style in addition to notation. From an array of four explanations, the student is asked to choose the one that best explains what is wrong with the performance he is hearing. The "faultiness" of any given interpretation may be due to lack of observance of dynamic levels and changes, incorrect articulation style or pattern, choice of a tempo not in agreement with the tempo marking, or excessive (or insufficient) rubato. Examples of FI items may be viewed in Appendix A.

Questions might be raised regarding the testing of the recognition of faulty interpretation because interpretation is likely to be rather subjective and personal. The researcher shares Hoffren's view that there are certain broad limits to acceptable interpretation. Teachers are expected to guide the interpretation of their students along culturally sanctioned lines.[20] When the music clearly indicates certain guidelines regarding tempo, dynamics, or articulation, there are deviations possible to an extent which could be classified, albeit subjectively, as faulty interpretation.

---

[19] Hereafter referred to as the FI group.

[20] James Hoffren, "A Test of Musical Expression," Council for Research in Music Education, Bulletin No. 2 (Winter, 1964), 32.

Historical classification. A total of sixty-seven excerpts from recordings are in the historical classification group.[21] In one version, the student is asked to indicate which one of four given years is the best estimate of the year of composition of the excerpt he is hearing.[22] In the other version, the terms Baroque, Classical, Romantic, and Modern are used in lieu of years;[23] other examples are included in Appendix A.

Popular classification. The popular classification group[24] requires the student to classify the excerpt he hears as being representative of acid rock, soul, folk, country-western, pop standard, "bubble gum", or folk rock styles.

Broad categories. When the proposal was written, three broad categories of items were proposed: pitch, rhythm, and interpretation. The category of style was added after submission of the proposal. Item construction, when concluded, yielded five groups in the pitch category (MI, HI, TC, ON, and EN), three groups in the rhythm category (RD, ORI, and IMS), two groups in the interpretation category (BP and FI) and two groups in the style category (HC and PC).

---

[21] Hereafter referred to as the HC group.

[22] Hereafter referred to as the HC(Y) group.

[23] Hereafter referred to as the HC(L) group.

[24] Hereafter referred to as the PC group.

PREPARATION FOR EMPIRICAL TRIALS AND PROGRAMMING

Cards. Item construction was, at first, largely conceptual. All materials had been selected and the content of the test stimuli determined, but it was believed necessary to have separate, discrete records of the test stimuli. The test questions with their answer arrays, the contents of the tapes in notation, and the content of notational displays were placed on 5 x 8 cards. This lengthy quasi-clerical process was justified because it would facilitate recording and manipulation of item order.

Recording. With the exception of the IMS group, all item groups required aural stimuli. The HC and PC excerpts were made via a Bogen model B61 phonograph on a Wollensak model T-1980 tape recorder. The other items were recorded using an Electrovoice dynamic cardioid microphone, model 676, and a Wollensak model T-1980 tape recorder. Scotch 175 tape was used. All aural stimuli were recorded monaurally on the left channel. The right channel was kept clear for the future addition of segments of 400 hz tone; these tones function as signals to the computer in the audio assembly process that is part of the construction of software for the IBM 1500 Instructional System utilized in this study.

The order of items within each group was randomized with the aid of random number tables.[25] Tape recordings were made at the convenience of the performers; i.e., all the clarinet excerpts were recorded together, all the piano excerpts were recorded together, etc. It was necessary to arrange the tapes into the proper random order through extensive splicing.

---

[25] Jerome C. R. Li, Statistical Inference, I (Ann Arbor, Michigan: Edwards Brothers, Inc., 1964), pp. 589-598.

EMPIRICAL TRIALS OF TEST ITEMS

Necessity to Establish Item Difficulty Indices

Arrangement of the items within each section into a scale ordered according to item difficulty was necessary to provide the bases for the sequential or incremental aspects of the test. If item $n + 1$ is more difficult than item n, the assumption can be made, theoretically, that the student who answers item $n + 1$ correctly will also answer item n correctly. Conversely, the student who is unable to answer item n correctly may be assumed unable to answer item $n + 1$ correctly. Since the test under development was planned to be incremental, i.e., every student would not receive every test item, such assumptions were necessary for a scoring procedure.

A conscious effort was made to vary the difficulty of items within each section. Range, instrumentation, and apparent conspicuity of the error were manipulated. Nevertheless, the difficulties of the completed items were unknown. Any attempt to order items according to difficulty would have been made on the basis of the researcher's personal estimate of item difficulty figures. Therefore it was necessary to administer each potential test item to undergraduate music education students to obtain an empirical estimate of item difficulty.

Preparation of Paper-and-Pencil Forms

A separate set of paper-and-pencil forms was prepared for each test section. Included in a set of forms were the response forms and, when necessary, notation sheets containing the notated musical examples to which the students were to listen. Conventional ditto masters, a typewriter, and a ballpoint pen were utilized. The staff lines were placed on a blank master with a typewriter. Notation was drawn freehand, with the aid of an ordinary ruler. All alphameric material, other than tempo markings, dynamic markings, and meter signatures, was typed. With the exception of the EN notation sheets, the end products were considered legible and adequate for the empirical trials.

Administration of Items

A total of 920 test items was constructed. The number of items made it impossible to administer each item at The Pennsylvania State University in the course of one term of ten weeks duration. A total of thirteen discrete periods of time, one period per test section, would have been an unreasonable disruption of normal instructional activity in music education classes, so thirteen other Pennsylvania institutions offering an undergraduate curriculum in music education were contacted and requested to provide time and students.

Of the thirteen institutions, six were able to offer the desired assistance, including Westminster College (New Wilmington), Carlow College (Pittsburgh), Bucknell University (Lewisburg), Susquehanna University (Selinsgrove), Temple University (Philadelphia), and Mansfield State College (Mansfield). Items were administered at those six institutions and at The Pennsylvania State University. Because of the difficulty of making scheduling arrangements, it was not possible to conclude the empirical administration of test items in the desired ten weeks; rather, it took approximately four months.

The propriety of establishing item difficulty indices at institutions other than The Pennsylvania State University, the institution for which the computerized test was being developed, may be questioned. If the item difficulties established as a result of testing at other institutions were grossly divergent from item difficulties that would have been established at Penn State, the scaling of items according to difficulty could lead to highly undesirable results. A strong difference in the relative ordering of items administered to Penn State students and administration to students elsewhere would be particularly disconcerting. This problem, however, was partially alleviated by calculating coefficients of rank-order correlation between the two orders of difficulty obtained for any subtest administered at different institutions. Highly significant coefficients (ρ's $\geq .85$) were interpreted as being indicative of necessary amount of consistency in difficulty rankings between two groups.

Administrations were conducted from the end of January to the end of April, 1970. In each case test forms and, when necessary, notation sheets were distributed. Tape recordings were played on a Wollensak T-1980 machine through the machine's internal speakers. The same machine was used at all locations. Each test form had a code number. Each student, identifying himself only by the code number of his test form, completed a data card by providing information regarding his institution, class standing, principal performing background, and curriculum. The purpose of the test was explained to the students; the point was stressed that the test itself, rather than the students, was being tested.

Melodic intervals. The MI test was administered to twelve students at The Pennsylvania State University on April 24, 1970 and to twenty-four students at Temple University on April 27, 1970. Each interval was played twice in anticipation of a repeat option that would be programmed into the computerized version of the final test. The tone quality of the tape appeared adequate for the purpose. Students at each location tended to feel that the MI test was rather easy; this was eventually supported by item difficulty data which showed a sparsity of difficult ($p \leq .30$)[26] items. Perhaps there would have been more difficult items if some intervals had been presented in descending order.

Harmonic intervals. On April 24, 1970 the HI test was administered to twelve music education students at The Pennsylvania State University. Twenty-one Temple University students had the test administered to them on April 27, 1970. Each interval was played twice. The HI test was apparently considerably more difficult than the MI test; there was a sparsity of easy ($p \geq .70$) items.

---

[26] A proportion of students equal to or less than .30 answered the item correctly. Item difficulty figures throughout this research were computed, in the conventional manner, by dividing the number of correct answers to each item by the number of students attempting each item. See G. P. Helmstadter, Principles of Psychological Measurement (New York: Appleton-Century-Crofts, 1964), p. 163.

Triad classification. Twenty-three undergraduate music education students at The Pennsylvania State University received the TC test on February 20, 1970; another twenty-one students at Susquehanna University received the test on February 26, 1970. Each of the forty-eight triads was repeated once. The tone quality of the tape was generally satisfactory, but the less-than-perfect ensemble of the amateur performance caused some distraction. Some of the more difficult triads were made more difficult by recording them at close spacing with combinations of instruments such as horn, trombone, and tuba. These combinations were occasionally found to be annoying to students. Perhaps the instrumentation occasionally made some triads, although legitimate, unrealistic in the context of traditional homophonic music.

Omitted notes. Two groups, one consisting of thirteen students and the other of nine students, were administered the ON test in a morning and afternoon session at Westminster College on January 26, 1970. The ON test was also administered to twenty-eight students at Carlow College on February 16, 1970. As in the other tests in the broad area of pitch, the ON test was administered with each tape recorded item being played twice. The quality of the notation sheets and the tape recordings appeared quite adequate for the purpose. Most students seemed to feel that it was unnecessary to repeat each item, but they welcomed the repetition of the more difficult items.

Erroneous notes. The EN test was not successful. It was administered to fifteen undergraduates in music education at Bucknell University on February 23, 1970. Fifteen students were considered to be an inadequate sample for the purpose of establishing item difficulty indices; the administration of the EN test was never repeated because the tape and, to an extent, the notation sheets were not adequate. All EN items are chorales, and they were performed by a pianist, a woodwind group, and a brass group. In spite of extensive recording sessions, the ensemble performances, particularly those prepared by the brass group, were inadequate. Error detection was further complicated by the sheer length of the test; it probably would have been better to have constructed fewer EN items. It was believed that the time necessary to revise the EN test could be spent more profitably with other tests.

Rhythmic discrepancies. Thirty students at The Pennsylvania State University received the RD test on February 5, 1970. The tape and notation sheets were adequate, but there was a problem caused by unintentional prompting. The student's task in the RD test is to follow the notation and indicate the number of the measure where what he hears is in rhythmic disagreement with what he sees. Since there is only one answer, once a student detects a discrepancy he can immediately indicate the measure. During the administration on February 6, a few students tended to respond because other students did; if a pencil moved during measure *n* of the performance, other pencils automatically followed. Instructions should have been given to wait until the music stopped before answering the item. Of course this would not be a problem in the final computerized, individualized version, but some results of the empirical trial may have been contaminated. Again, it was believed to be better to spend on another test the time needed for retrial of the RD test.

Overall rhythmic inaccuracies. Temple University was the site of the administration of the ORI test on April 27, 1970. The administration appeared to go smoothly; directions were clear, and tapes and notation sheets were adequate. There were no complaints from the twenty-five students regarding the amount of time required to answer the questions or the nature of the questions.

Better phrasing. The BP test was administered twice. On February 27, 1970, it was administered to twenty students at The Pennsylvania State University; the second administration was to ten students at Mansfield State College on April 28, 1970. The notation sheets and tapes were adequate.

Faulty interpretation. No unforeseen problems occurred during administration of the FI test on February 20, 1970 and February 26, 1970 to twenty-two students at The Pennsylvania State University and to twenty-one students at Susquehanna University, respectively. The issue of subjectivity was not raised by the students; there appeared to be ample time to answer the questions. Quality of the notation sheets and the sound reproduction were adequate for the purpose.

Historical classification. The HC(Y) version was administered at Carlow College on February 16, 1970. The thirty-one students generally enjoyed the test; there were no difficulties with the test materials.

The HC(L) version was administered to twenty-six students at Susquehanna University on February 27, 1970 and to thirteen students at The Pennsylvania State University on April 3, 1970. There were no difficulties.

The HC(L) version of the HC test asks the student to choose a letter to indicate his classification of each musical excerpt as representative of the Baroque, Classical, Romantic, or Modern Period. The HC(Y) version asks the student to choose from an array of four years the one he believes is the most likely date of the excerpt's composition. When the results of the empirical trials were examined and twenty-item scales were selected from the HC(Y) and HC(L) item pools, it was found that substantially different items were selected. Items that were relatively difficult in one version were relatively simple in the other version. It may be possible to conclude that students have processes for classifying excerpts by years that are different from their processes for classifying identical excerpts by musical periods. The HC(L) version was chosen for future use as a HC test.

Nonadministered tests. Time became a crucial factor; two tests were never administered. The PC test was developed after consultation with an experienced radio and television man, but the categories of acid rock, soul, folk, country-western, pop standard, "bubble gum," and folk rock may not be ample. Rock music is often difficult to classify into a discrete category; many examples are "hybrids" - stylistic indicators of two or more styles may be present. The Music Educators Journal's extensive treatment of youth music[27] suggests that perhaps the PC test is in need of some conceptual revision prior to any administration.

---

[27] Music Educators National Conference, "Youth Music - A Special Report," Music Educators Journal, LVI (November, 1969), 43-74.

The IMS test was also unadministered. Unlike the other tests, the IMS test contains no aural stimuli. Consequently, as time became crucial, it was given a lower priority than the other rhythm area tests.

Postadministration Analysis

Data analysis. Details of the data analysis will be reported in the succeeding chapter. An item difficulty index was computed for each item by dividing the number of correct responses to each item by the number of respondents attempting the item.

Using the difficulty indices as a guide, a twenty-item scale was selected from the pool of items for each test. The responses given by each student who participated in the testing sessions were written as a series of coded answer strings, one string per student. Then, a hypothetical answer string was written for each student, based upon the responses the student gave to items that would have been presented to the student in accordance with the programming strategy had the student taken the test through the IBM 1500 Instructional System. Items that would not have been presented in the computerized version were coded as incorrect responses if they were higher in the scale (i.e., closer to item 20) than the highest presented item answered correctly. Items not presented that were lower in the scale than the highest presented item answered correctly were coded as correct responses. Each student's string of actual correct and incorrect responses to the selected items for each test was compared with the hypothetical string of responses that would have resulted from the student answering identically the items presented through a computerized version of the test.

The comparison of answer strings served as a basis for the computation of descriptive statistics which showed in various ways relationships between the empirical trial and proposed computerized versions of the test segments. A product-moment correlation coefficient showed the size and degree of relationship between the actual number of correct responses on the selected twenty-item scale for each student and the hypothetical number of correct responses that would have been attributed to each student based upon the programming strategy. An "accuracy" figure was computed by subtracting the number of mispredictions of student responses resulting from the programming strategy divided by the number of possible predictions from 1.00. A correlated $\underline{t}$ test was applied to the distribution of $\underline{N}$ difference scores, i.e., the actual number of correct responses subtracted from the hypothetical number of correct responses for each student on each twenty-item scale. The null hypothesis was that the mean of the actual-hypothetical differences was not significantly different from zero. A rank-order correlation figure was computed for twenty-item scales selected from tests which were administered at more than one campus to show the relationship of item difficulties at the two locations. These data will be reported in the following chapter.

Selection of tests for programming. Nine tests were developed and administered to samples large enough to provide meaningful data, but the number of tests to be programmed was limited to four. The amount of time expended on the item development and empirical trial stages was far greater than originally anticipated. Furthermore, a test limited to four sections would fit concisely into the 75-minute class period at The Pennsylvania State University, and more detailed analysis could be done with fewer tests.

The original commitment, through the proposal funded by the U. S. Office of Education, was to develop a prototype computerized, criterion-referenced test which would purport to measure certain nonperformance musical behaviors in the broad areas of pitch, rhythm, and interpretation. The area of style was added to the overall design after submission of the proposal. It appeared logical that the tests selected for programming should represent each area.

The ON test was selected to represent the pitch area. It seemed to be the most musically interesting of the pitch tests because the items were melodies rather than isolated tonal stimuli.

The ORI test was selected to represent the rhythm area. Of the two rhythm tests that were administered, the ORI test appeared to have the greater strength: The scale of difficulties yielded more nearly equal intervals.

The FI test was selected to represent the interpretation area. Taking the test seemed to require a broader range of thinking than the BP test, and the empirical trials of the FI test had been quite satisfactory.

The HC test, in the HC(L) version, was selected to represent the style area. The HC(L) version was the one that had been successfully administered to students at The Pennsylvania State University; the low rank-order correlation of difficulty rankings ($\rho = .53$) between the HC(L) and HC(Y) twenty-item scales indicated, in part, that the two forms of the HC test were rather different.

For a more detailed explanation of the selection process, the reader is referred to Appendix F.

PROGRAMMING THE COMPUTERIZED TEST

IBM 1500 Instructional System

The medium for presenting the computerized test was the IBM 1500 Instructional System, housed in the Computer-Assisted Instruction Laboratory of The Pennsylvania State University. The self-contained system, operational at Penn State since January 1968, is designed for individualized instruction; its capacity for rapid access and coordination of stimuli and rapid processing of student responses makes the system useful for testing.

Central to the 1500 System is the IBM 1131 Central Processing Unit which provides active storage for all system data. A vast amount of additional data may be brought into the central processing unit from disk cartridges mounted on IBM 2310 Disk Storage Drives. In addition to controlling the processing of data, the central processing unit controls the physical operation of the other components of the IBM 1500 Instructional System, including a card read punch, a printer, and the components of the student instructional stations.

The student instructional stations, also referred to as terminals or stations, consist of a cathode ray tube screen (CRT), a typewriter keyboard, an image projector, a light pen, and an audio unit. The conventional arrangement of the instructional station places the CRT mounted atop the typewriter directly in front of the seated student. The image projector is to the left of the CRT; the light pen is to its right. The audio unit is above the CRT.

The CRT resembles a television screen. Sixteen horizontal rows and forty vertical columns may be coordinated to provide a total of 640 positions in which aplhameric characters or special symbols, such as musical notation, may be displayed. Characters most frequently appear on the screen as white on a dark blue background. Test questions and answer areas for the test reported herein are always displayed on the CRT.

Students taking the test answer questions by firmly pressing the light pen to a lighted area on the CRT coded to the answer of their choice. The light pen receives light from the screen and transmits the location of the student response to the system which then takes the action for which it has been programmed, e.g., scoring a response.

Although the typewriter may be used for input of student responses, in the current test the typewriter is used only for initial student contact with the computer ("signing on") and occasionally changing the display on the CRT.

The image projector, containing a 7.5 by 9-inch screen on which photographic images may be shown, is used for all displays of musical notation. Image cartridges containing 16mm film may contain as many as 1,000 discrete photographs. The system has the capacity to access individual image frames at the rate of 40 frames per second; therefore, any particular combination of notational displays could be arranged in a desired program sequence with no necessary consideration of image access.

Headphones connected to the audio unit are used to present aural stimuli. Tape cartridges mounted in audio units may contain as many as two hours of taped messages. The four-track tape used in the cartridges contains three message tracks and one digital signal address track to allow the location of any particular message.

The Coursewriter II programming system is used with the IBM 1500 Instructional System. The author of material to be presented through the system writes coded instructions in the Coursewriter language to direct the presentation of content to the student. Material to be printed on the CRT and its location, segments of tape to be played, action to be taken in the event of specific student responses, and what image to show must be programmed into the computer. An example of Coursewriter programming from the computerized test may be viewed in Appendix D.

Programming Strategy

One principal characteristic of the computerized test of certain nonperformance musical behaviors is its incrementalization. Originally, a (→5), (←3), (→2), (←1), (→1) strategy was proposed; that is, the student would start with the fifth item in a series of twenty. A correct response would branch him ahead to the tenth item (an increment of five), but an incorrect response would branch him back to the second item (a reverse increment of three). After one error, the forward increment, following a correct response, would be two. Occurrence of a second error would branch the student back one item and change the forward increment to one; a third error would terminate the administration of the test section.

During the analysis of data obtained from empirical trials of test items, it was apparent that the original strategy would tend to cause premature terminations for some students. A straight linear strategy, in which every student would receive every item, would result in no mispredictions but would be inefficient and, to a computer programmer, conceptually alarming. A modified linear strategy was adopted, in place of the original strategy, as a compromise between duplication of off-line results and efficiency in amount of items presented. Under the modified linear strategy, a student starts with the fourth item in a twenty-item scale. He continues to receive items in increments of four as long as he emits no incorrect response. The first error causes a reverse branch of three and changes the forward increment to one. The student then continues ahead regardless of the correctness of a response until he makes a total of five errors or three successive errors.

Scoring Procedure

Originally, the number of the most difficult item answered correctly was planned to be the tested student's earned score. Considerable study of student answer strings revealed that somewhat spurious conclusions could result in instances where a student might fail to answer numerous items but nevertheless manage to answer correctly one item of high difficulty. Therefore, rather than using scale scores, each student's score for each of the four programmed tests was expressed simply in terms of the number of items answered correctly. The student who answered more items correctly than another student probably progressed further along the scale; he had fewer strings of consecutive incorrect answers.

Audio Preparation

All test items had been recorded prior to empirical trials of test items. From conventional tape recordings, audio cartridges for the IBM 1500 Instructional System must be prepared through a special process.

The musical excerpts for the selected items were spliced into the item order for the final computerized version. Using the IBM cue tone generator and a Roberts model 1040 tape recorder, 400 hz tone segments were then placed on the right channels of the tapes. These 400 hz cue tones functioned as signals to the computer during the audio assembly process; breaks in the continuity of the 400 hz tone indicated the end of one tape message (i.e., musical excerpt) and the beginning of another.

After the original tapes contained the cue tone, the audio assembly process was activated. The tapes were mounted on an Ampex special model tape recorder with remote control capacity. An IBM four-track tape cartridge was mounted in the audio unit at one of the instructional stations. A special computer program was utilized to duplicate each message and assign to each message a unique digital address, thereby permitting the accessing of any particular musical excerpt by the Coursewriter program.

The master tape cartridge produced during the audio assembly process was duplicated with a Viking model 235 tape duplicator to produce the tape cartridges used in the administration of the test.

Film Preparation

Film preparation included preparation of the art work, photography, and film processing. The only stage with which the researcher was directly involved was the preparation of the art work, i.e., notation sheets. Each musical example was copied with a black felt tip pen on to white paper ruled with staff lines. The quality of manuscript notation was judged to be quite adequate for the purpose.

After photographing of the notation sheets, the film was processed through the regular channels utilized by the Penn State CAI Laboratory for the preparation of film cartridges. Five cartridges were made; each cartridge contained one exposure of each image, identified with a digital address to permit access in the Coursewriter program.

Debugging

Extensive examination and trial of the Coursewriter program was conducted by the researcher to detect and remove faulty coding (i.e., "bugs") from the program. Grammatical errors such as invalid codes and erroneous parameters are of relatively little concern with the Coursewriter programming system because the computer will not accept statements containing such errors. Subtle errors in programming can result from simple typographical errors, however; results quite different from those anticipated can be obtained because of a programmer's momentary lapses in accuracy. For example, during the debugging process, it was discovered that the score for the Historical Classification section was often inaccurate; the score indicated by the computer did not reflect the total number of correct responses accredited to the student. Investigation located an error in the programming segment specifying action to be taken in the event of a correct response to the fourth item in the HC scale, the item initially presented to the student. The student was intended to receive four points since the assumption was made that items one, two, and three could have been answered correctly if item four was answered correctly. But an instruction that should have said ad 4>/c4 read ad 1>/c4. A simple mistake in numerals caused inaccurate scoring.

FINAL ADMINISTRATION

Student Population

The computerized test was administered to thirty-two students during the week of October 5-9, 1970. A parallel conventional version was administered during the same week to twenty-eight students. All students were undergraduate music education majors at The Pennsylvania State University.

It was considered desirable to look for gross differences in scores between upper-term and lower-term students[28] because, if the criteria upon which test items were based are representative of competency development currently transpiring at Penn State, there should be such differences. (Lack of such differences could be attributed to lack of sensitivity in the test as well as lack of representativeness in the criteria.) All first, second, third, and fourth term students (N = 36) were chosen to participate in the study, as well as all eighth, ninth, tenth, eleventh, twelfth, and over-twelfth term students (N = 36). Each student was randomly assigned to either the computerized version or the

---

[28] The traditional terms "freshmen," "sophomores," "juniors," and "seniors" are rarely used at Penn State. The University academic year is divided into four ten-week terms; an undergraduate student is classified on the basis of his term standing. Since undergraduates in music education generally require twelve terms to complete their degree requirements, students classified as first, second, or third term could be called "freshmen," students classified as fourth, fifth, or sixth term students could be called "sophomores," etc.

O

E

1

F

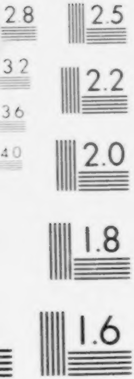
D

25

1.0
1.1
1.25
1.4
4.5
5.0
5.6

MICROCOPY RESOLUTI

NATIONAL BUREAU OF STA


2.8
2.5
3.2
2.2
3.6
2.0
4.0
1.8
1.6

TEST CHART

parallel conventional version, so that there were eighteen upper-term and eighteen lower-term students assigned to each testing situation. The desired number was fifteen students per term grouping per testing situation; the excess was to allow for loss of a few students.

Administrative Procedure, Computerized Version

Students assigned to the computerized version were assigned a time to report to the Computer-Assisted Instruction Laboratory during the week of October 5-9, 1970. Upon arrival for his testing session, each student was assured by the researcher that the test rather than the student was being tested. Operation of the light pen was explained, and each student was shown how to adjust the volume of the audio unit output. The student was assured that the researcher would be available if needed, the door to the testing room was closed, and the test program was permitted to run its course. At the conclusion of the test, the student's four subtest scores were automatically output by a typewriter connected to the computer, and the researcher asked the student for an opinion.

The items administered to each student were determined, in accordance with the programming strategy, by the response history of the student. The student was permitted to repeat a taped excerpt for any ON item once if he wished; the other items were played only once. If a student did not respond to any item within forty-five seconds to the end of the taped excerpt, that was considered to be an incorrect response.

Administrative Procedure, Conventional Version

Students assigned to the conventional testing condition were asked to report to a central location on October 6, 1970. The test which these students received was similar to the tests utilized for the earlier empirical trials. Each student used a mimeographed test form containing printed instructions and eighty test items identical to the items comprising the four twenty-item scales programmed for the computerized version. The necessary notation for each item appeared on mimeographed notation sheets. The original tapes were duplicated; these duplicates were then edited to provide approximately eight seconds of silence between examples in the ON and HC sections, and approximately twenty seconds of silence between items in the ORI and FI sections.

The researcher administered the test. Students were assured that the test was being tested, rather than they. Tape-recorded instructions supplemented printed instructions; students were permitted to ask questions. All ON items were repeated; other items were played once.

Plan for Analysis of Data

A questionnaire was appended to each test. Each student was asked which section of the test was the most difficult and the least difficult for him. He was asked whether, if he had a choice, he would have preferred to take the computerized or conventional versions. He was also asked to evaluate the quality of sound reproduction and notation as well as the amount of pressure he felt while taking the test.

A procedure outlined by Medley[29] was utilized to investigate the equivalency of the two versions of the test. According to Medley, two tests are equivalent only if four stringent criteria are satisfied. The students must be ranked in the same order by the two tests, the variances of errors of measurement must be equal, the variances of test scores must be equal, and the test means must be equal. These conditions are tested by means of F tests after analysis of variance summary tables, similar to those suggested by Hoyt for estimating test reliability in terms of internal consistency,[30] have been plotted. The Medley procedure was utilized because it might indicate the divergence of the computerized test from the conventional version, or, in gross terms, what price one must pay in terms of differing results for the convenience of computerized testing of this nature.

SUMMARY STATEMENT OF METHOD OF TEST DEVELOPMENT

The initial stage of test development was to frame a series of objectives which could be used as criteria upon which to build a criterion-referenced test. Test items were constructed in relation to those criteria. After empirical trial of test items, certain items were scaled according to difficulty, and four tests were selected for final administration. The computerized test and a parallel conventional test were administered to undergraduate music education majors, and the resulting data were analyzed.

---

[29] Donald M. Medley, "A General Procedure for Testing the Equivalence of Two Tests" (paper read at meeting of the National Council on Measurement Usage in Education, February 19, 1957, New York).

[30] Cyril Hoyt, "Test Reliability Obtained by Analysis of Variance," Psychometrika, VI (June, 1941), 1953-160.

CHAPTER IV

RESULTS AND FINDINGS

The purpose of this chapter is to present and interpret data from the item trial, selection, and final administration stages of the computerized and conventional versions of the test. The general procedure will be to discuss the purpose of the particular data collection and processing, present the data, and offer an interpretation of it.

PRELIMINARY DATA

Preliminary data include data gathered regarding test items prior to the final administration of the test. Item difficulty indices and data resulting from comparison between results from actually administering selected items to students and results from hypothetically administering items to students in accordance with a programming strategy are included. Such data are reported herein to aid the reader's understanding of the processes of development.

Computation of Item Difficulty Indices

After administration of a section or subtest to a group of undergraduate students majoring in music education, the items comprising that section were scored. The item difficulty index for each item was computed by dividing the number of correct responses to an item by the number of students attempting the item.[1] This was done for each section.

---

[1] When a student failed to respond, his lack of response was nevertheless considered to be an incorrect response and an "attempt."

Item difficulty indices obtained ranged from 1.00 (all students responded correctly to the item) to 0.00 (no students responded correctly to the item). It might have been desirable to obtain item difficulties in approximately equal numbers at equi-incremental points along the range (e.g., three items with ID = .95, four items with ID = .90, three items with ID = .85, . . ., four items with ID = .05), but based upon the empirical trials, items tended to cluster more toward the less difficult end of the scale.

A twenty-item scale was selected for each subtest administered to twenty-five or more students. The primary criterion for selection of an item was the difficulty index; when more then one item was available for selection at a given level of difficulty, selection was also based upon musical criteria such as the quality of the performance.

Table 1 shows the item difficulty indices for the twenty items selected for each subtest. It may be noted that the greatest amount of difference between any two adjacent items is .26; the least amount of difference is .00.

Actual-Hypothetical Comparisons

After selection of items for twenty-item scales, answer strings were written for each student to whom the subtest had been administered. An answer string consisted of a string of 1's, indicating correct response, and 0's, indicating incorrect responses. For example, here is the answer string for one student's responses to the twenty-item ON scale:

1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 0 0 0 0 0

Table 1

Item Difficulty Indices of Selected
Items for Twenty-item Scales

Test Section*a	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
ON	.98	.96	.94	.92	.88	.84	.80	.76	.72	.68	.64	.60	.56	.52	.48	.42	.36	.30	.24	.18
ORI	1.00	.96	.92	.88	.84	.80	.76	.72	.68	.64	.60	.56	.52	.48	.40	.32	.28	.20	.12	.04
FI	.98	.95	.93	.91	.88	.84	.77	.72	.67	.63	.58	.53	.49	.44	.40	.35	.38	.23	.19	.16
HC(L)	1.00	.97	.92	.87	.82	.77	.74	.69	.64	.59	.54	.49	.44	.38	.36	.33	.26	.21	.12	.08
HC(Y)	1.00	.97	.94	.87	.84	.81	.77	.74	.71	.68	.65	.58	.52	.45	.39	.35	.29	.23	.16	.10
MI	1.00	.97	.94	.92	.89	.86	.83	.81	.78	.75	.72	.69	.67	.64	.61	.58	.53	.44	.33	.25
HI	.97	.94	.91	.88	.82	.76	.73	.70	.67	.61	.55	.48	.42	.36	.30	.24	.18	.12	.06	.00
TC	.98	.95	.91	.86	.82	.77	.73	.70	.68	.64	.61	.59	.55	.52	.48	.43	.39	.32	.23	.18
RD	1.00	.97	.93	.90	.87	.83	.80	.77	.73	.70	.63	.60	.57	.50	.47	.43	.17	.03	.03	.00
BP	1.00	.97	.93	.90	.87	.83	.80	.77	.73	.70	.63	.57	.50	.43	.37	.30	.23	.17	.10	.03

---

*a For identity of the abbreviations and descriptions of the test sections the reader is referred to Chapter III, pages 33-38.

This particular student was able to answer the first twelve items in the scale correctly.[2] After that, he was able to answer only the fifteenth item correctly.

For mathematical convenience, the assumption was made that a student would respond to an identical item in an identical manner although the mode of presentation was different. This was believed to be a conservative assumption because it denied the researcher the opportunity to expect nonequivalent responses and thus account for unexpected variance. If the items coded in the above answer string were presented to the same student through the IBM 1500 Instructional System, the student, if he behaved in accordance with the assumption, would again answer the first twelve items correctly, answer the next two incorrectly, correctly answer the fifteenth item, and miss the remaining five items.

Once the assumption of equivalent responses to identical items was made, it was possible to construct hypothetical answer strings to represent a student's responses in accordance with a programming strategy. Here is a comparison between the hypothetical answer string for the above student, in accordance with the programming strategy eventually adopted, and the actual answer string that resulted from the empirical trials of the ON items:

Hypothetical: 1 1 1 $\underline{1}$ 1 1 1 $\underline{1}$ 1 1 1 $\underline{1}$ $\underline{0}$ $\underline{0}$ 0 $\underline{0}$ 0 0 0 0

Actual: 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 0 0 0 0 0

---

[2] In the empirical trials conducted to obtain the item difficulty indices, the order of item presentation was determined with the aid of a random number table. Hence, the order of presentation of the twenty items eventually choosen to comprise the scale was not, at the time of the trials, 1, 2, 3, . . ., 20.

The underlined numerals indicate items that would have been presented in the computerized version. The nonunderlined numerals in the hypothetical string indicate items for which a correct (1) or incorrect (0) response was assumed. In this case, the student would have been presented with six items and earned a score of twelve. His correct answer to the fourth item, his initial item, would have branched him to the eighth item. The correct answer to the eighth and then to the twelfth item would have continued the increment of four. The incorrect response to the sixteenth item would have caused a reverse branch to item thirteen and changed the forward increment to one. Items thirteen and fourteen would have been answered incorrectly; under the assumption, the three successive errors (sixteen, thirteen, fourteen) would have terminated the ON test for this student.

From a series of comparisons between answer strings, it was possible to compute various descriptive statistics. One statistic upon which importance was placed by the researcher was the correlation between the actual scores of students for each twenty-item scale and the hypothetical scores that would have resulted from a computerized version. The original programming strategy was abandoned, in part, because the revised strategy adopted raised these correlations. These figures are reported in Table 2.

Responses to items that would not have been administered to a student were assumed to be correct if they were to items of less difficulty than the last item administered, and assumed to be incorrect if they were to items of greater difficulty than the last item administered.

Table 2

Descriptive Statistics Based upon Comparisons of Hypothetical and Actual Answer Strings

Test Section	Number to Whom Administered	r Between Scores	Prediction Accuracy	*t* Test
ON	50	.89	.85	NS
ORI	25	.85	.84	NS
FI	43	.82	.82	NS
HC(L)	39	.83	.83	NS
HC(Y)	31	.87	.83	$p < .10$
MI	36	.88	.84	$p < .10$
HI	33	.93	.88	NS
TC	44	.90	.81	NS
RD	30	.84	.83	NS
BP	30	.79	.84	NS

The following is a hypothetical answer string that would have resulted, under the assumption of equivalent responses, from a student receiving the ORI scale in accordance with the adopted programming strategy:

$$1\ 1\ 1\ \underline{1}\ 1\ 1\ 1\ \underline{1}\ \underline{0}\ \underline{1}\ \underline{0}\ \underline{0}\ \underline{1}\ \underline{1}\ \underline{0}\ \underline{0}\ 0\ 0\ 0\ 0$$

Here, the student was hypothetically administered ten items, those items for which the response codes are underlined. The response codes for nonadministered items are assumed to be 1 (correct) if they are of less difficulty than the last item administered and 0 (incorrect) if they are of greater difficulty than the last item administered, item sixteen. The ten codes for nonadministered items may be said to represent predictions of responses.

Consideration of the same hypothetical ORI answer string when it is matched with the actual answer string yields the following:

Hypothetical: $1\ 1\ 1\ \underline{1}\ 1\ 1\ 1\ \underline{1}\ \underline{0}\ \underline{1}\ \underline{0}\ \underline{0}\ \underline{1}\ \underline{1}\ \underline{0}\ \underline{0}\ 0\ 0\ 0\ 0$

Actual: $1\ 1\ 1\ 1\ 0\ 0\ 1\ 1\ 0\ 1\ 0\ 0\ 1\ 1\ 0\ 0\ 1\ 0\ 0\ 0$

Of the ten predicted responses, it is apparent that there were mispredictions for items five, six, and seventeen. The remaining seven predictions were accurate. The quantity of mispredictions for a given student could vary from zero to twenty minus the number of items administered; in mathematical language,

$$0 \leq M \leq (20 - A),$$

where M indicates the number of mispredictions for a given student and A indicates the number of items hypothetically administered to that student. By summing the number of mispredictions across all students, dividing that sum by the quantity obtained from subtracting the total number of items hypothetically administered from the total number of students multiplied by twenty (i.e., the total number of predictions), and subtracting the quotient from 1.00 it is possible to obtain an index of prediction accuracy. The formula for the index of prediction accuracy for a subtest may be written as

$$P = 1.00 - \frac{\Sigma M}{20n - \Sigma A},$$

where P represents the index of prediction accuracy, M represents the number of mispredictions for a student, A represents the number of items hypothetically administered to a student, and n represents the number of students to whom the subtest was administered. Indices of prediction accuracy are reported in Table 2.

When the students' actual scores for twenty-item scales were matched with their hypothetical scores, a series of difference scores (hypothetical minus actual scores) was computed. The aim was to have essentially the same scores result from hypothetical and actual versions. A null hypothesis was formulated to state that there was no difference between the mean of the difference scores and zero. A correlated t test was applied for each subtest; as Table 2 indicates, the t values were nonsignificant except for the MI and HC(Y) tests.

The data in Table 2 were based upon the assumption of response equivalency. To the extent that the assumption was valid, the data were a valid means of evaluating the tests which were constructed. It must be noted, however, that the data do not attempt to describe a relationship between an actual administration and a hypothetical administration to different students.

Administration at Varying Institutions

It was impossible to administer all sections of the test under development to music education undergraduate students at The Pennsylvania State University because of the constraints of time. Consequently, as described in Chapter III, empirical trials of test items were also conducted at six other Pennsylvania institutions of higher education which offer an undergraduate curriculum in music education. When a test was administered at more than one institution, a rank-order coefficient of correlation was computed to show the relationship between the two sets of rankings (in terms of item difficulty) assigned to the items chosen to comprise a twenty-item scale. A low rank-order correlation coefficient (ρ) would indicate considerable diversity in difficulty order of the items. Seven test sections were administered at more than one institution; the number of students tested and their division by institutions as well as the computed ρ for each test are contained in Table 3.

Examination of Table 3 reveals that four of the seven tests administered at more than one institution yielded a ρ greater than .85. Two test sections were in the range .70 - .85; the MI test was below .70. All are significant beyond the .001 level when one uses the modified t test for significance of rank-order correlation suggested by Bruning and Kintz,[3] but $\rho \approx .90$ was considered more desirable than $\approx .70$.

---

[3] James L. Bruning and B. L. Kintz, Computational Handbook of Statistics (Glenview, Illinois: Scott, Foresman and Company, 1968), pp. 158-159.

Table 3

Rank-order Coefficients of Correlation and Number of Students Tested per Institution for Tests Administered at More Than One Institution

Test Section	Number Tested	Institutional Division	ρ
ON	50	Carlow College, 28; Westminster College, 22	.93
FI	43	Penn State University, 22; Susquehanna University, 21	.94
HC(L)	39	Penn State University, 13; Susquehanna University, 26	.89
MI	36	Penn State University, 12; Temple University, 24	.69
HI	33	Penn State University, 12; Temple University, 21	.95
TC	44	Penn State University, 23; Susquehanna University, 21	.82
BP	30	Penn State University, 20; Mansfield State College, 10	.80

DATA FROM FINAL ADMINISTRATION

The final administration[4] occurred October 5-9, 1970. The computerized test and the parallel conventional test were compared through the Medley procedure, discussed earlier. Comparisons between the scores of lower-term and upper-term students were made; responses to a questionnaire appended to both versions were studied.

Medley Procedure

The Medley procedure is illustrated through Table 4 which summarizes the procedure for the ON test administered to the total number of students (thirty-two in the computerized version, twenty-eight in the conventional version). An analysis of variance was performed for the group that received the computerized version, the group that received the conventional version, and the combined groups. These analyses of variance partitioned the total variance into variance attributable to differences among students, differences among item means, and error. The sums of squares (SS) for the components of variance were computed, as Medley suggested, in accordance with Hoyt's formulas,

$$\text{SS among students} = \frac{1}{n} \Sigma t_k^2 - \frac{(\Sigma t_k)^2}{nk},$$

$$\text{SS among items} = \frac{1}{k} \Sigma p_n^2 - \frac{(\Sigma t_k)^2}{nk},$$

and

$$\text{total SS} = \frac{(\Sigma t_k)\ (nk - \Sigma t_k)}{nk},$$

---

[4] The term "final administration" means final with regard to the research reported herein. The reader should not conclude that computerized testing of nonperformance musical behaviors has had its final hour.

Table 4

Medley Procedure for ON Test, All Students

Source of Variation	DF	SS	MS
A. Group receiving computerized version			
(1) Students	31	29.9984	0.9677
(2) Items	19	57.4984	3.0262
(3) Error	589	62.2516	0.1057
(4) Total	639	149.7484	
B. Group receiving conventional version			
(5) Students	27	16.8314	0.6230
(6) Items	19	27.0785	1.4252
(7) Error	513	80.3215	0.1566
(8) Total	559	124.2214	
C. Combined groups			
(9) Students	59	47.3292	0.8022
(10) Items	19	77.9292	4.1015
(11) Error	1121	149.2208	0.1331
(12) Total	1191	274.4792	
D. Analysis of equivalence			
(13) Groups (9 - 1 - 5)	1	0.5094	0.5094
(14) Students (1 + 5)	58	46.8198	0.8072
(15) Items (10)	19	77.9292	4.1015
(16) Error between versions (11 - 3 - 7)	19	6.6477	0.3499
(17) Error within versions (3 + 7)	1102	142.5731	0.1294
(18) Total (12)	1199	274.4792	

Test for Criterion 1: Are students ranked in same order?

$\underline{F} = \frac{MS(16)}{MS(17)} = 2.7040$; p <.005, criterion not met

Test for Criterion 2: Are variances of errors of measurement equal?

$\underline{F} = \frac{MS(7)}{MS(3)} = 1.4816$; p <.005, criterion not met

Test for Criterion 3: Are variances of obtained scores equal?

$\underline{F} = \frac{MS(1)}{MS(5)} = 1.5533$; NS, criterion met.

Test for Criterion 4: Are means equal?

$\underline{F} = \frac{MS(14)}{MS(13)} = 1.5846$; NS, criterion met.

where n represents the number of items (twenty in Table 4), k represents the number of students (thirty-two, twenty-eight, and sixty for groups A, B, and C, respectively, in Table 4), t_k represents any particular student's score, and p_n represents the particular number of correct responses to any particular item.[5] The data necessary for use of the Hoyt formulas were readily obtainable from the typewritten score summary and student records provided by the computer for the computerized version or the test papers for the conventional version.

After partitioning of the variance into components for each testing group and the combined group, the analysis of equivalence was made. Section D of Table 4 includes the quantities, indicated in parentheses after the names of the sources of variation, which were added or subtracted, in accordance with the Medley procedure, to obtain the degrees of freedom and SS figures for Section D. For example, the degrees of freedom and SS for students were found by adding the appropriate quantities for (1), variation attributable to students who received the computerized version and (5), variation attributable to students who received the conventional version.

Mean squares (MS), obtained by dividing SS by the appropriate degrees of freedom, provided the needed quantities for the four F tests used to test the four criteria for equivalence. Criterion one, ranking of students in the same order by each version of the test, or homogeneity of function, was tested by comparing $MS_{(16)}$ with $MS_{(17)}$. For the ON test as it was administered to all students, the F value obtained in testing for criterion one is significant beyond the .005 level; it can be said that there is no difference in the ranking of students yielded by the two versions must be rejected, and criterion one is not satisfied. Criterion two, equality of variances of errors of measurement, was tested by comparing $MS_{(3)}$ with $MS_{(7)}$; in the case of the ON test, illustrated by Table 4, this criterion was also not met. Criterion three, equality of variance of obtained scores from the two versions, was met; it was tested by comparing $MS_{(1)}$ with $MS_{(5)}$, and the obtained F value was not significant. Criterion four, equality of means, was tested by comparing $MS_{(13)}$ with $MS_{(14)}$; the ON test evidently met this criterion.

---

[5] Cyril Hoyt, "Test Reliability Obtained by Analysis of Variance," Psychometrika, VI (June, 1941), 154.

Summary tables, similar to Table 4, will be found in Appendix B for applications of the Medley procedure to the four programmed tests for the total number of students, the lower-term students only, and the upper-term students only. Table 5, a summary of all the applications, indicates that no test met all criteria; YES indicates a non-significant F value, and NO indicates a significant F value. The equality of means criterion was most frequently met; only the HC test failed. The other criteria were met either rarely or never.

No section of the computerized test may be said to be equivalent to its corresponding conventional section. The process of computerization with its incremental feature may be said to have distorted the test beyond the point of equivalency. But what is the practical meaning of the lack of equivalency?

To fulfill criterion one, both versions of the test should rank the students in the same order. Item differences should interact no more with differences among one group of students than with differences among

Table 5

Summary of Medley Procedure Applications, Indicating Presence of Equivalence of Tests According to Four Criteria

Test	Criterion 1: Are Students ranked in same order?	Criterion 2: Are Variances of errors of measurement equal?	Criterion 3: Are Variances of obtained scores equal?	Criterion 4: Are Means equal?
A. For all students				
ON	NO	NO	YES	YES
ORI	NO	NO	NO	YES
FI	NO	NO	NO	YES
HC	YES	NO	NO	YES
B. For lower-term students only				
ON	NO	NO	YES	YES
ORI	NO	NO	NO	YES
FI	NO	NO	NO	YES
HC	NO	NO	NO	NO
C. For upper-term students only				
ON	NO	NO	YES	YES
ORI	YES	NO	NO	YES
FI	NO	NO	NO	YES
HC	NO	NO	YES	YES

another group of students. But if such interaction does differ, as it did in ten of twelve cases, how critical are the differences? The difference in the ranking effect of identical items in the two versions may have been attributable to different rank orders in terms of item difficulty, as will be presented below. Since the purpose of the criterion-referenced test was not to rank students, criterion one may have less significance for a criterion-referenced test than for a norm-referenced test.

Fulfillment of criterion two requires equality of the errors of measurement which occur in any measurement situation. The assumption of responses to nonadministered items based upon responses to administered items in the computerized version introduced systematic error to the extent that the assumed rank of the nonadministered items in terms of difficulty differed from their actual rank. The complete lack of attainment of criterion two is one serious flaw in the test as it was administered.

Equality of variances of obtained scores, criterion three, occurred only for the ON test for the three groupings of students, and for the HC test for upper-term students. Failure to meet this criterion may, again, be traced to inaccurate positions of items in the twenty-item scales. Difficult items toward the supposedly easy end of a scale could have caused premature terminations of a computerized test section; easy items toward the supposedly difficult end would not have been reached by terminated students but would have been presented to students who received the conventional version.

Equality of means occurred for all test sections except the HC test. In considering the two modes of test presentation, equality of means might lead to the conclusion that, on the average, the differing test versions would have given identical scores and facilitated identical interpretations of those scores with regard to what, if any, action should be instigated as a result of the scores. However, the computerized score of a particular student might not be representative of his status regarding the musical behaviors being measured. Equality of means accompanied by nonequality of variances of obtained scores may have resulted from a balance between students who received the computerized version and were terminated prematurely with students who spuriously received credit for correct responses to nonadministered items. Again, this is related to the discrepancy between presumed rankings of item difficulty and actual rankings in the testing situation.

All Medley criteria call for comparisons of variances which should lack statistically significant differences. The researcher believes that the significant differences observed are related to the divergency between expected and actual rankings of test items in terms of difficulty.

Comparison of Item Difficulty Rankings

Empirical trials were conducted to establish item difficulty indices. Items were selected to form twenty-item scales for each test section which was administered to at least twenty-five students; the difficulty indices for selected items are reported in Table 1 above. The strategy was to develop tests in twelve areas related to nonperformance musical behaviors. (Concern for the refinement of programming strategy and the constraints of time were responsible for the reduction of the number of test sections programmed to four.) Hence, many test items were administered to small groups of students during empirical trials to establish item difficulty indices. The instability of those indices may have been responsible for the quantitative difficulties with the test.

Table 6 contains the estimated item difficulty indices for the four twenty-item scales; the estimates are, of course, the difficulties obtained from the empirical trials. The observed difficulties for the computerized version and the conventional version,[6] computed in the usual manner, are also contained in the table. Item difficulty figures for the computerized test are partially based on assumed responses. Discrepancies occur in certain instances, for example, the eighteenth ORI item, the sixth FI item, and the fourteenth HC item. Some items, of course, such as the fourth ON item and the seventeenth ORI item have very similar figures.

The rank order of item difficulties varies from scale to scale. Ideally, the coefficient of rank-order correlation RHO (ρ) should be 1.00 between any two sets of item difficulty indices for one test section. Rank-order correlations are reported in Table 7; the correlation between the estimated difficulty indices and the observed indices from administration of the conventional version varies from .43 to .87.

Less than perfect rank order of item difficulties means that for the computerized version students received credit for nonadministered

---

[6] Item difficulties are reported on the basis of administration to the total number of students taking each version because there was no distinction between students regarding class standing during the empirical trials.

Table 6

Discrepancies Between Estimated and Observed Item Difficulties

Scale	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
ON est.	.98	.96	.94	.92	.88	.84	.80	.76	.72	.68	.64	.60	.56	.52	.48	.42	.36	.30	.24	.18
ON obs. (comp.)	1.00	1.00	1.00	.91	.88	.97	.81	.66	.69	.81	.75	.59	.50	.44	.50	.41	.28	.12	.16	.09
ON obs. (conv.)	1.00	.89	.89	.89	.82	.79	.75	.68	.79	.64	.93	.68	.61	.46	.46	.71	.50	.39	.18	.29
ORI est.	1.00	.96	.92	.88	.84	.80	.76	.72	.68	.64	.60	.56	.52	.48	.40	.32	.28	.20	.12	.04
ORI obs. (comp.)	1.00	.97	.94	.78	.91	.84	.84	.7	.62	.62	.75	.62	.56	.31	.41	.19	.28	.12	.06	.03
ORI obs. (conv.)	.89	.93	.71	.82	.71	.61	.82	.8	.57	.79	.75	.46	.57	.57	.57	.29	.25	.36	.07	.29
FI est.	.98	.95	.93	.91	.88	.84	.77	.72	.67	.63	.58	.53	.49	.44	.40	.35	.28	.23	.19	.16
FI obs. (comp.)	.97	1.00	.94	.84	.97	.72	.75	.62	.78	.78	.66	.59	.66	.44	.41	.44	.25	.03	.03	.00
FI obs. (conv.)	1.00	.93	.68	.71	.82	.57	.61	.82	.64	.71	.50	.61	.68	.75	.54	.68	.54	.25	.36	.25
HC est.	1.00	.97	.92	.87	.82	.77	.74	.69	.64	.59	.54	.49	.44	.38	.36	.33	.26	.21	.13	.08
HC obs. (comp.)	.97	1.00	.97	.50	.56	.66	.38	.44	.44	.47	.38	.34	.28	.12	.16	.12	.12	.16	.12	.00
HC obs. (conv.)	1.00	.96	.79	.50	.43	.68	.21	.75	.46	.61	.43	.50	.57	.71	.50	.39	.43	.75	.54	.07

Table 7

Rank-order Coefficients of Correlation for Difficulty Rankings

Test	Estimated and Computerized Observed	Estimated and Conventional Observed	Computerized and Conventional Observed
ON	.98	.87	.88
ORI	.98	.87	.86
FI	.95	.74	.75
HC	.96	.43	.54

items of a difficulty level greater than those administered items which were answered correctly to permit that credit for nonadministered items. For example, if a student answered the first two items presented in the computerized ORI test section, items four and eight, correctly, he earned eight points and was ready for item twelve, but item three, according to the difficulty estimate from the conventional version, was more difficult than item four, and items five, six, and seven were more difficult than item eight. Assuming that the item difficulty indices computed from administration of the conventional version were accurate estimates of the difficulty of the items for those who received the computerized version, nonincrementalization (i.e., administering all items in the computerized version to all students in a linear manner) would have made possible a greater degree of equivalence.

Comparison of Test Performance of Upper-term and Lower-term Students

If the skills measured by the ON, ORI, FI, and HC tests are increased during the undergraduate training of the music education student at The Pennsylvania State University, the mean performance of the upper-term students should have been greater than the mean performance of the lower-term students. Greater upper-term mean scores could indicate that what was tested was pertinent to the present focus of the curriculum.

Table 8 reveals that, with one exception, the mean score for lower-term students was always lower than the mean score for upper-term students; however, in only one instance was the difference statistically significant according to a *t* test. Upper-term students differed only

Table 8

Comparison of Upper-term and Lower-term Mean Scores

Test Section	Test Version	Lower-Term Mean	Upper-Term Mean	_t_-test
ON	Computerized	10.83	14.71	$p < .05$
ON	Conventional	13.38	13.33	NS
ORI	Computerized	11.00	12.50	NS
ORI	Conventional	11.31	12.67	NS
FI	Computerized	10.94	13.07	NS
FI	Conventional	12.44	12.92	NS
HC	Computerized	7.17	10.07	NS
HC	Conventional	10.62	12.25	NS

slightly from lower-term students in their mean ability to identify notes missing from a passage, detect an accurate explanation of rhythmic inaccuracy, choose an explanation of the departure from tasteful interpretation, and classify musical examples by periods of music history.

Failure to find greater differences between the mean test scores of upper-term and lower-term students may be attributed to a possible lack of curricular experience directed toward improvement of the skills measured. It may also be attributed to a possible lack of relevancy to present coursework on the part of the test; however, it was not intended to develop the test within the confines of the present course structure.

Questionnaire Results

A seven-item questionnaire was appended to each test version. Student opinion was sought regarding relative difficulty of the test sections, quality of sound and notation, speededness of the test, pressure placed on the student, and preferred version. Students who received the computerized version answered the multiple-choice questions with the light pen; students who received the conventional version checked their responses. All students in each group answered each question with one response only. Tables 9 through 15 summarize the questionnaire responses in terms of proportions of the students indicating each response. The questionnaire items are presented in Appendix C.

There was no particular expectancy regarding the test sections considered the most or the least difficult. These findings are reported in Tables 9 and 10. In each case the trend is more clear for the

Table 9

Questionnaire Responses Regarding Most Difficult Section

Section Indicated	Proportion Choosing Response, Computerized Version	Proportion Choosing Response, Conventional Version
ON	.125	.000
ORI	.219	.286
FI	.375	.179
HC	.281	.536

Table 10

Questionnaire Responses Regarding Least Difficult Section

Section Indicated	Proportion Choosing Response, Computerized Version	Proportion Choosing Response Conventional Version
ON	.312	.714
ORI	.250	.071
FI	.094	.071
HC	.344	.143

conventional than for the computerized version. The students who received the computerized version did not know the number of items in each section, since the amount of items presented to any student varied with the student's performance in accordance with the incremental programming strategy. All students who received the conventional version received each test item and thereby had a greater number of items upon which to base a decision regarding difficulty. Neither test version gave knowledge of results to any student prior to administration of the questionnaire; no student's estimation of section difficulty was influenced by any knowledge of his relative success among the sections.

The quality of sound reproduction in the computerized version of the test was of concern. The IBM 1506 audio unit, the tape playback component of the IBM 1500 Instructional System, always contained white noise, a constant background hissing sound, while musical excerpts were played. Deihl noted this hissing sound as well as bubbling sounds, apparently caused by momentary disruption of the uniform movement of the tape during a stage of audio cartridge preparation, and variance in sound quality between tracks of the tape.[7] These unmusical qualities, plus occasional static, raised the possibility that students might find certain items difficult to answer for an extraneous reason.

It was expected, therefore, that students who received the conventional version of the test, with its tape recorded at 7.5 ips, one generation removed from the original recordings, would evaluate the sound quality to be at a higher level than the students who received the computerized test, with its tape recorded at 1.875 ips, two generations removed from the original recordings, because of the extraneous noise on the 1506 tape cartridge. But, as indicated in Table 11, a greater proportion of students to whom the computerized version was administered chose the most favorable response. This was not expected by the researcher; perhaps students, while they listened for relevant cues with which to select an answer, were more oblivious to extraneous noise in an individualized situation, or perhaps the wearing of headphones had some influence.

---

[7] Ned C. Deihl, Development and Evaluation of Computer-Assisted Instruction in Instrumental Music, Project No. 7-0760, ERIC No. ED 035 314. (Washington: Office of Education, U. S. Department of Health, Education, and Welfare, 1969), p. 36.

Table 12 summarizes the questionnaire responses regarding the quality of the notation. It was expected that few students receiving the computerized version would find the professionally processed film exposures of painstakingly drawn music manuscript to be of low quality. The mimeographed notation sheets used by the students who received the conventional version of the test were certainly not illegible, but were not comparable to printed music.

Perceived speededness of the test versions was of interest. The medium of computer-assisted instruction appears to lend itself well to individualization of presentation; rates of presentation of material can be varied greatly to accomodate students of varying work habits and abilities. It is possible to program presentations for student control; the material appearing on the cathode ray tube need not change until the change is requested by the student. Unlimited allowances for time to respond are not considered desirable in the computerized test under discussion, but a full forty-five seconds is allowed between the time the playback of a musical excerpt concludes and the time the student is

Table 11

Questionnaire Responses Regarding Quality of Sound Reproduction

Response	Proportion Choosing Response, Computerized Version	Proportion Choosing Response, Conventional Version
Very poor and distracting; it made the questions difficult to answer.	.031	.036
Not good, but it did not interfere with my ability to answer the questions.	.094	.179
Fair, it certainly was adequate for the test.	.500	.536
Quite good; it was often enjoyable to listen	.375	.250

Table 12

Questionnaire Responses Regarding Quality of Notation

Response	Proportion Choosing Response, Computerized Version	Proportion Choosing Response, Conventional Version
Very poor; the illegibility of the notes often made it difficult to answer questions.	.031	.000
Not good; but it did not interfere with my ability to answer the questions.	.031	.321
Not comparable to printed music, but it was certainly adequate for the purpose.	.344	.464
Quite good; it was comparable to printed music in most respects.	.594	.214

automatically considered to not know the answer.[8] Forty-five seconds was believed to be sufficient for virtually any student, but the time allotments for the conventional version--eight, twenty, twenty, and eight seconds respectively for the ON, ORI, FI, and HC tests--were planned with an average student in mind.

It was expected that most students who received the computerized version would find that their test moved at a comfortable pace while more than a few students who received the conventional version would find that their test moved either too slowly or too rapidly. The expected results were partially found; as Table 13 indicates, most students found the speed of the computerized version to be satisfactory. It was, however, interesting that more students did not find the conventional version to be too rapid.

Assurances were given to all students in each group that the test, not the student, was being tested. Nevertheless, the researcher was interested in obtaining some indication of tension or pressure felt by the students. Unfamiliarity with computers and other electronic apparatus might have been conducive to an increase in tension; mere placement in a testing situation, in spite of assurances given to the student, might have increased tension. Table 14 summarizes the questionnaire data regarding perceived tension; it is apparent that the very few instances of more than slight tension which occurred were in the group who received the conventional version. No particular result was anticipated.

---

[8] In the case of the ON test, where students have the option of repeating an excerpt once before responding, the forty-five seconds are counted in full from the time of conclusion of the second play.

Table 13

Questionnaire Responses Regarding Speededness of the Test

Response	Proportion Choosing Response, Computerized Version	Proportion Choosing Response, Conventional Version
Too slowly; there was needless delay between items.	.031	.036
At a comfortable pace for me.	.938	.607
Too rapidly; there was insufficient time between items.	.031	.357

Table 14

Questionnaire Responses Regarding Perceived Pressure and Tension

Response	Proportion Choosing Response, Computerized Version	Proportion Choosing Response, Conventional Version
Quite calm and relaxed; there was very little pressure on me.	.594	.464
Slightly tense; there was some pressure on me, but it was largely of my own making.	.406	.464
Rather tense; pressure was being placed upon me by the testing situation.	.000	.071
Quite tense and agitated; I was constantly being pressured and urged to produce answers.	.000	.000

The final item in the questionnaire asked the student whether, if there had been a choice, he would have preferred one version of the test over the other version. All students knew prior to the testing dates that a computerized version of an experimental test in music was to be compared with a conventional version of the same test. No student, of course, received both versions; the students were asked to express a preference after being familiarized with only one version. Expectation was that the majority response for each group would be that it made no difference which version the student received while slightly more than half of the remaining responses from each group would indicate a preference for the familiar version. Within the group receiving the computerized version, the proportion expressing preference for their version was the majority. The proportions of responses among the group receiving the conventional version were in accordance with expectation. Table 15 summmarizes the preference data.

NON-QUANTITATIVE FINDINGS

The IBM 1500 Instructional System functioned smoothly and efficiently during all stages of test development. Malfunctions within the program were always found to be the result of human error. In all instances, a student who was scheduled to be tested could report to the Penn State Computer-Assisted Instruction Laboratory, have the operation of equipment briefly explained to him, and begin the test within two minutes of his arrival.

Table 15

Questionnaire Responses Regarding Preference of Testing Situation

Response	Proportion Choosing Response, Computerized Version	Proportion Choosing Response, Conventional Version
The CAI Laboratory, using the computerized instructional station in an individualized manner.	.656	.179
A conventional paper-and-pencil testing situation, as part of a group taking the test simultaneously.	.031	.214
It really made no difference.	.312	.607

There were no problems related to any slowness of the system during administration of the test. An excessive amount of input from other stations can slow the presentation of material to a student at a given instructional station, particularly when the input is an addition or replacement of coded instruction. This did not occur.[9]

Numerous students commented that their experience with the computerized instructional station was novel, enjoyable, or worthy of replication. There was no apparent apprehension regarding the equipment. One student stated a desire that all of his tests could be administered in the same manner.

The brief summary of scores printed at a typewriter station by the computer at the conclusion of each testing session was always rapidly available in the following format:

```
STUDENT     x17
ON score is   8
ORI score is  4
FI score is   7
HC score is   3
```

If the tests were refined to the point where some action could be taken on the basis of the scores, the quick score summary would be very beneficial.

The extensive student records available from the computer provide an accurate record of each student's testing session. Information contained in student records includes question identifiers, response identifiers, student identifiers, time of response, and time elapsed between the end of musical excerpt and entry of response. From the student records it was easy to obtain data for item analysis and determine which items were actually administered to any student. A sample of student records is presented in Appendix E.

---

[9] Had it occurred, it could have been alleviated by restricting system usage during testing sessions to execution of existing programmed material rather than creation or alteration of material.

CHAPTER V

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

This final chapter summarizes the conduct of the research and the findings, states conclusions, and presents some recommendations for further research.

SUMMARY

Objectives

The framing of valid objectives upon which to build criterion-referenced test items was the initial phase of the research. Objectives were stated in the form of observable nonperformance musical behaviors. Quantitative statements were avoided; objectives were statements of skills which were deemed important for display by competent music education graduates. Areas included by the objectives, not intended to be an all-inclusive statement of desirable nonperformance musical behaviors, were:

aural recognition and identification of melodic intervals;

aural recognition and identification of harmonic intervals;

aural recognition and classification of triads;

insertion of missing notes into visual notational displays of aurally perceived melodies;

recognition and location of aural-visual pitch discrepancies in four-part harmonic passages;

recognition and location of aural-visual rhythmic discrepancies;

selection of appropriate explanations of incorrectly performed rhythmic patterns;

recognition and location of incorrectly notated measures for given meter signatures;

selection of members of pairs of examples that are performed "better" when "better" refers to tapered phrase endings, dynamics, appropriateness of breathing, or appropriateness of articulation style;

identification and classification of inappropriateness of interpretation when the inappropriateness is due to inappropriate tempo, inappropriate articulation, excessive rubato, lack of rubato, or inappropriate dynamics;

classification of musical examples as being stylistically representative of the Baroque, Classical, Romantic, or Modern Period;

classification of musical examples as being stylistically representative of acid rock, soul, country-western, pop standard, "bubble gum," folk, folk rock, or blues.

Test Items

Multiple-choice items were constructed in accordance with the above objectives using orchestral excerpts, chorales, and pedagogical literature. Items were notated, recorded, and prepared for empirical trial to establish item difficulty indices.

Nine test sections were subjected to empirical trials at The Pennsylvania State University and six other Pennsylvania institutions of higher education offering an undergraduate music education curriculum. Twenty-item scales arranged in order of difficulty were selected from each test section. Actual student performances on those scales were compared with hypothetical performances which would have resulted from equivalent responses to those scales as they would have been presented through a proposed programming strategy. On the basis of the empirical trials and descriptive statistics obtained from the actual-hypothetical comparisons, and in consideration of the four basic areas of pitch, rhythm, interpretation, and style, four test sections were selected for programming. The selected test sections were the Omitted Notes, Overall Rhythmic Inaccuracies, Faulty Interpretations, and Historical Classification sections, related to the fourth, seventh, tenth, and eleventh of the objectives summarized above.

Programming

The selected items were programmed in the Coursewriter II language for the IBM 1500 Instructional System. An incremental programming strategy was utilized; a student began each computerized test section with the fourth item of the twenty-item scale. A correct response branched the student to the eighth item; the student continued to move ahead in increments of four items until an initial erroneous response occurred or the twentieth item was answered correctly. An initial erroneous response caused a reverse branch of three items; e.g., if a student was unable to answer item twelve correctly, he was branched to item nine. From the point reached by the reverse branch after the initial erroneous response, the student moved ahead along the scale in a linear manner. A test section was terminated for a student when he reached the end of the scale, made three erroneous responses in succession, or made a total of five erroneous responses. His score was the number of test items actually answered correctly plus the number of test items assumed to be answered correctly. Nonadministered items were assumed to be answered correctly if they were lower on the scale (i.e., were of less difficulty) than the highest administered item on the scale that was answered correctly.

Administration and Findings

The computerized test was administered to eighteen lower-term and fourteen upper-term undergraduates enrolled in the music education curriculum at The Pennsylvania State University during the week of October 5-9, 1970. A parallel conventional version of the test was administered to sixteen lower-term and twelve upper-term students to provide a check on the item difficulties and a basis for a comparison of test equivalence. Students who received the computerized version worked at an instructional station; they heard the musical stimuli through headphones, read the test questions on the cathode ray tube screen, viewed musical notation on the image projector, and answered questions by indicating their choices with a light pen. Students who received the conventional version were seated in a classroom; they read the questions and answered them on mimeographed test forms, viewed musical notation on mimeographed notation sheets, and heard the musical excerpts through the speakers of the tape recorder.

The Medley procedure, a series of F tests for equal variances, was utilized to test for equivalence of the two versions in accordance with four criteria: Equal ranking of students, equality of variances of errors of measurement, equality of variances of obtained scores, and equality of means. Although the equality of means criterion was generally met, the others were not; the two versions of the test may not be considered equivalent.

Neither the computerized nor the conventional version of the test showed any significant difference between the mean scores of upper-term and lower-term students. It was not clear that this was a weakness of the test because the students' curricular experiences may not be directed toward improvement of the skills measured.

The weakness of the test, preventing its immediate implementation, is the discrepancy between the estimated item difficulty indices, established as a result of the empirical trials of test items, and the actual item difficulty indices, computed from the conventional version scores. This discrepancy caused assumptions regarding correctness of nonadministered items in the computerized version to be less than accurate.

The computerized test was well received by the students to whom it was administered. The equipment functioned smoothly, and audio weaknesses present in the IBM 1500 Instructional System did not appear to have any adverse effect upon the test.

CONCLUSIONS

Four conclusions may be drawn from the present study:

1. Present skills, techniques, and equipment are adequate for the construction of a workable computerized criterion-referenced test of certain nonperformance musical behaviors.

2. Rank order of items, in terms of item difficulty, is critical to the success of an incremental programming strategy in computerized testing wherein assumptions are to be made regarding responses to nonadministered items.

3. The computerized criterion-referenced test of certain nonperformance musical behaviors is not equivalent to a conventional noncomputerized version of the test.

4. Differentiation of mean scores between lower-term and upper-term students is minor and generally nonsignificant; it is uncertain as to whether this is a function of the test or lack of significant growth in the skills measured.

RECOMMENDATIONS

Further research is recommended to refine the computerized test and increase its potential utility for The Pennsylvania State University and its paradigmatic value for other institutions. More accurate item difficulty indices are required; perhaps the empirical establishment of such figures could be preceded or supplemented by rational study of the musical behaviors involved. Additional objectives should probably be formulated and new types of test items constructed from them. Alternate programming strategies might be actually programmed and compared. Specifically, the following recommendations are made:

1. Existing test items should be administered to large groups (N = 200) of undergraduate music education majors in order to obtain more accurate estimates of item difficulties.

2. In some cases, the grouping of test items according to difficulty might be approached by analyzing the behaviors involved in responding to the items and establishing an ordered series of prerequisite behaviors.

3. Additional objectives related to nonperformance musical behaviors should be formulated and test items constructed; however, this should not precede the strengthening of existent items.

4. After the reordering of test items on the basis of stronger estimates of difficulty, a three-group study should be conducted to compare the relative merits of 1) a computerized test programmed in a manner identical to the test developed in the study reported herein, 2) a computerized test programmed following a differing strategy, and 3) a parallel conventional version of the test.

At the beginning of the first chapter it was stated that the basic purpose of the study was to develop a prototype computerized criterion-referenced test for measuring competencies in certain nonperformance musical behaviors present in undergraduate students commencing their course of study in music education. The prototype has been largely developed. If the recommendations can be implemented, a new and useful instrument will exist.

BIBLIOGRAPHY

BOOKS AND PUBLISHED TESTS

Aliferis, James. Aliferis Music Achievement Test (College Entrance Level). Minneapolis, Minnesota: University of Minnesota Press, 1954.

Bruning, James L., and B. L. Kintz. Computational Handbook of Statistics. Glenview, Illinois: Scott, Foresman and Company, 1968.

Cronbach, Lee J. Essentials of Psychological Testing. 2d ed. New York: Harper and Row, Inc., 1960.

Farnsworth, Paul R. The Social Psychology of Music. New York: Holt, Rinehart, and Winston, Inc., 1958.

Gagné, Robert M. The Conditions of Learning. New York: Holt, Rinehart, and Winston, Inc., 1966.

Gordon, Edwin. Musical Aptitude Profile. Boston: Houghton-Mifflin Company, 1965.

Helmstadter, G. P. Principles of Psychological Measurement. New York: Appleton-Century-Crofts, 1964.

Hightower, Caroline. How Much are Students Learning? Plan for a National Assessment of Education. Ann Arbor, Michigan: The Committee on Assessing the Progress of Education, 1968.

Kibler, Robert J., Larry L. Barker, and David T. Miles. Behavioral Objectives and Instruction. Boston: Allyn and Bacon, Inc., 1970.

Lehman, Paul R. Tests and Measurements in Music. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1968.

Leonhard, Charles, and Robert W. House. Foundations and Principles of Music Education. New York: McGraw-Hill Book Company, Inc., 1959.

Li, Jerome C. R. Statistical Inference. 2 vols. Ann Arbor, Michigan: Edwards Brothers, Inc., 1964.

Lindvall, C. M. Measuring Pupil Achievement and Aptitude. New York: Harcourt, Brace, and World, Inc., 1967.

Lindvall, C. M. Testing and Evaluation: An Introduction. New York: Harcourt, Brace, and World, Inc., 1961.

Lundin, Robert W. An Objective Psychology of Music. 2d ed. New York: Ronald Press, 1967.

Whybrew, William K. Measurement and Evaluation in Music. Dubuque, Iowa: The William C. Brown Company, 1962.

PUBLICATIONS OF THE GOVERNMENT AND UNIVERSITIES

Deihl, Ned C. Development and Evaluation of Computer-Assisted Instruction in Instrumental Music. Project No. 7-0760, ERIC No. ED 035 314. Washington: Office of Education, U. S. Department of Health, Education, and Welfare, 1969.

French, Joseph L. "Numerical and Verbal Aptitude Tests Administered at the CAI Student Station," Semi-Annual Progress Report (prepared by Harold E. Mitzel et al), Experimentation with Computer-Assisted Instruction in Technical Education. Project No. 5-85-074. University Park, Pa.: The Pennsylvania State University Computer-Assisted Instruction Laboratory, 1967.

Glaser, Robert. Evaluation of Instruction and Changing Educational Models. C. S. E. I. P., Occasional Report No. 13. Los Angeles: University of California at Los Angeles Center for the Study of Evaluation and Instructional Programs, 1968.

Hansen, Duncan N. An Investigation of Computer-Based Science Testing. FSU CAI Center, Semiannual Progress Report, Report No. 6. (prepared by Duncan N. Hansen, Walter Dick, and Henry T. Lippert). Tallahassee, Florida: Florida State University Computer-Assisted Instruction Center, 1968.

Lippert, Henry R. and Walter Ehlers. Computer-Based Testing. FSU CAI Center, Annual Progress Report, Report No. 7 (prepared by Duncan N. Hansen, Walter Dick, and Henry T. Lippert). Tallahassee, Florida: Florida State University Computer-Assisted Instruction Center, 1968.

PERIODICAL LITERATURE

Ball, Charles Hershel. "The Application of an Empirical Method to the Construction of a College Entrance Test in Music." Unpublished doctoral dissertation, George Peabody College for Teachers, 1964. Dissertation Abstracts, XXVI (July-August, 1965), 404.

Bloom, Benjamin S. "Learning for Mastery." Evaluation Comment, I (May, 1968), 1-12.

Cox, Richard C. and Glenn T. Graham. "The Development of a Sequentially Scaled Achievement Test." Journal of Educational Measurement, III (Summer, 1966), 147-150.

Douglas, Charles Herbert. "Measuring and Equalizing Music Theory Competence of Freshman College Music Majors." Unpublished doctoral dissertation, The Florida State University, 1965. Dissertation Abstracts, XXVI (February, 1966), 4712.

Glaser, Robert. "Instructional Technology and the Measurement of Learning Outcomes: Some Questions," American Psychologist, XVIII (August, 1963), 519-521.

Gordon, Edwin. "Implications for the Use of the Musical Aptitude Profile with College and University Freshman Music Students," Journal of Research in Music Education, XV (Spring, 1967), 32-40.

Greer, Harry Holt, Jr. "The Application of a Digital Computer to Scoring and Analysis of Examinations and the Preparation of Diagnostic Reports." Unpublished docotral dissertation, The George Washington University, 1966. Dissertation Abstracts, XXVII (September-October, 1966), 923A.

Hatfield, Warren Gates. "An Investigation of the Diagnostic Validity of the Musical Aptitude Profile with Respect to Instrumental Music Performance." Unpublished doctoral dissertation, The University of Iowa, 1967. Dissertation Abstracts, XXVIII (January-February, 1968), 3210 A.

Hoffren, James. "A Test of Musical Expression," Council for Research in Music Education, Bulletin No. 2 (Winter, 1964), 32-35.

Hoyt, Cyril. "Test Reliability Obtained by Analysis of Variance," Psychometrika, VI (June, 1941), 153-160.

Jorgenson, James. "Advice to the Potential College Music Major," Instrumentalist, XXII (April, 1968), 38-39.

Lehman, Paul R. "A Selected Bibliography of Works on Music Testing," Journal of Research in Music Education, XVII (Winter, 1969), 428-442.

MENC Commission on Teacher Education. "Teacher Education in Music: An Interim Report of the MENC Commission on Teacher Education," Music Educators Journal, LVII (October, 1970), 33-48.

Mitzel, Harold E. "The IMPENDING Instruction Revolution," Phi Delta Kappan, LI (April, 1970), 434-439.

Music Educators National Conference. "Youth Music - A Special Report," Music Educators Journal, LVI (November, 1969), 43-74.

Perry, William Wade. "A Comparative Study of Selected Tests for Predicting Proficiency in Collegiate Music Theory." Unpublished doctoral dissertation, North Texas State University, 1965. Dissertation Abstracts, XXVI (January, 1966), 3995-3996.

Popham, W. James, and T. R. Husek. "Implications of Criterion-Referenced Measurement," Journal of Educational Measurement, VI (Spring, 1969), 1-9.

Williams, Gilbert. "The Use of the Computer for Testing, Programming, and Instruction," Research in Education, III (May, 1968), 105.

ARTICLES IN COLLECTIONS

Edling, Jack V. "New Media Applications," Man-Machine Systems in Education, John W. Laughary, editor. New York: Harper and Row, Inc., 1966. Pp. 69-79.

Glaser, Robert, and Richard C. Cox. "Criterion-Referenced Testing for the Measurement of Educational Outcomes," Instructional Process and Media Innovation, Robert A. Weisgerber, editor. Chicago: Rand McNally and Company, Inc., 1968. Pp. 545-550.

Glaser, Robert, and David J. Klaus. "Proficiency Measurement: Assessing Human Performance," Psychological Principles in System Development, Robert M. Gagné, editor. New York: Holt, Rinehart, and Winston, Inc., 1962. Pp. 418-474.

Tyler, Ralph W. "Changing Concepts of Educational Evaluation," Perspectives of Curriculum Evaluation, Ralph W. Tyler, Robert M. Gagné, and Michael Scriven, editors. Chicago: Rand McNally and Company, Inc., 1967. Pp. 13-18.

UNPUBLISHED MATERIALS

Cox, Richard C., and Julie S. Vargas. "A Comparison of Item Selection Techniques for Norm-referenced and Criterion-referenced Tests." Paper read at the annual meeting of the National Council on Measurement in Education, February, 1966, Chicago.

Mansur, Paul Max. "An Objective Performance-Related Music Achievement Test." Unpublished Doctor's dissertation, The University of Oklahoma, 1965.

Medley, Donald M. "A General Procedure for Testing the Equivalence of Two Tests." Paper read at the meeting of the National Council on Measurement Usage in Education, February 19, 1957, New York.

APPENDIX A

SAMPLE ITEMS

SAMPLE ITEMS

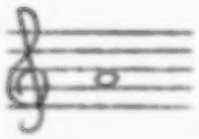
Two sample items are included from each of the four test sections which were programmed. The questions, the answer arrays, the notational displays, and the contents of the recorded excerpt are indicated for each item. Content of the item was identical for each version of the test. The reader will recall that in the computerized version, notational displays appeared on the image projector, questions and answer arrays appeared on the cathode ray tube, and the recorded music was heard through individual headphones. In the conventional version, the recorded music was played on a tape recorded for a group; the visual material was mimeographed.

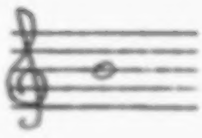
Omitted Notes, Item No. 5

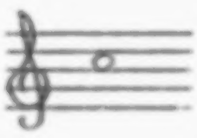
Question

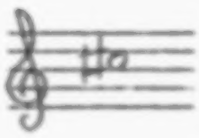
What is the name of the missing note?

Answer Array

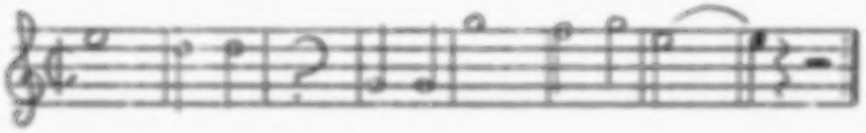




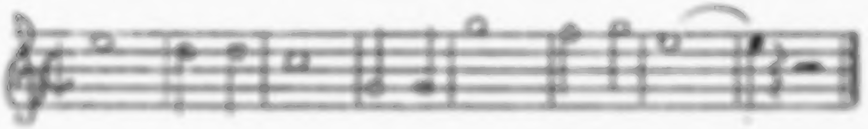




Notational Display



Contents of Recorded Excerpt



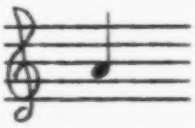
(Played on piano)

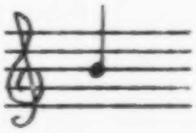
Omitted Notes, Item No. 14

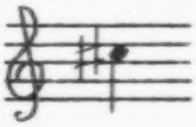
Question

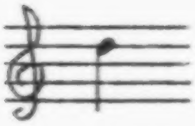
What is the name of the missing note?

Answer Array

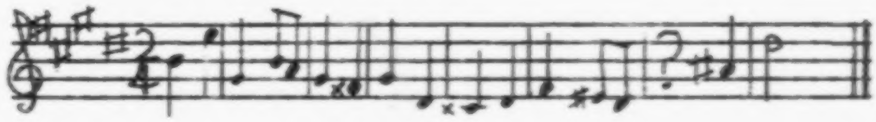




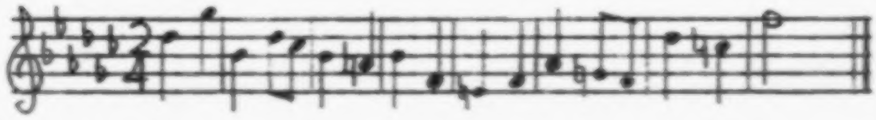




Notational Display



Contents of Recorded Excerpt



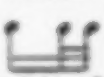
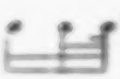
(Played on trumpet)

Overall Rhythmic Inaccuracies, Item No. 9

Question

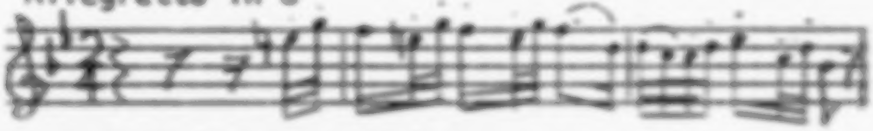
What is wrong with the rhythmic performance of this excerpt?

Answer Array

A. The tempo accelerates.
B. The thirty-second notes are played as sixty-fourth notes.
C. Unwritten ties are added.
D. The  groups are played as  groups.

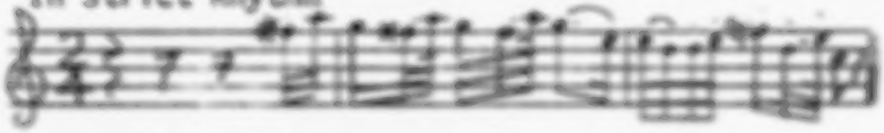
Notational Display

Allegretto in 8



Contents of Recorded Excerpt

In Strict Rhythm



(Played on clarinet)

Overall Rhythmic Inaccuracies, Item No. 20

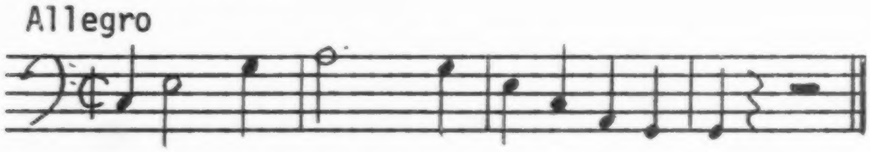
Questions

What is wrong with the rhythmic performance of this excerpt?

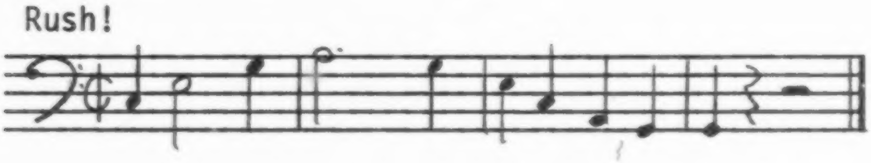
Answer Array

A. The tempo accelerates.
B. The tempo decelerates.
C. The quarter notes are played as half notes.
D. There is nothing wrong with the rhythmic performance.

Notational Display



Contents of Recorded Excerpt



(Played on euphonium)

Faulty Interpretations, Item No. 4

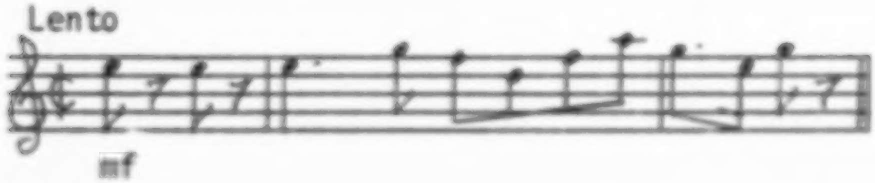
Question

What is wrong with the performer's interpretation of this melody?

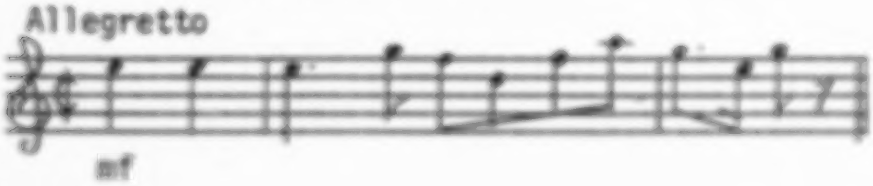
Answer Array

A. The rubato is excessive.
B. The rubato is insufficient.
C. The tempo is inappropriate.
D. The articulation is incorrect.

Notational Display



Contents of Recorded Excerpt



(Played on flute)

Faulty Interpretations, Item No. 11

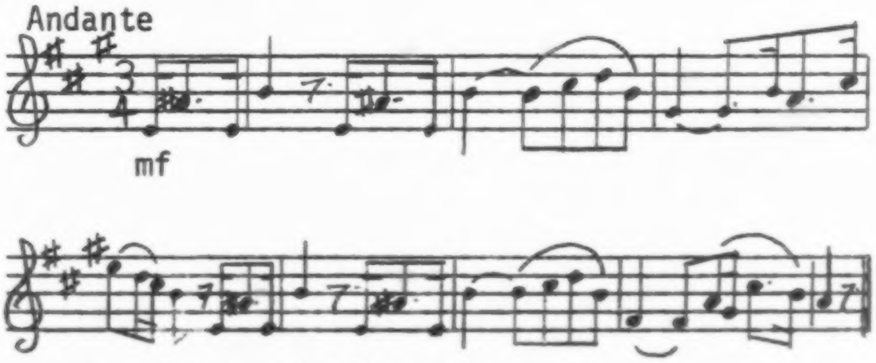
Question

What is wrong with the performer's interpretation of this melody?

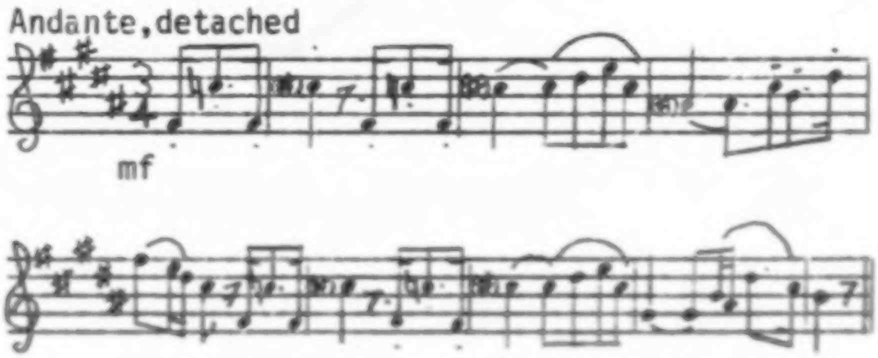
Answer Array

A. The tempo is inappropriate.
B. The articulation is incorrect.
C. The dynamics are unobserved.
D. The rubato is excessive.

Notational Display



Contents of Recorded Excerpt



(Played on trumpet)

Historical Classification, Item No. 2

Question

Is this excerpt most representative of the Baroque, Classical, Romantic, or Modern Period?

Answer Array

B C R M

Notational Display

(none)

Contents of Recorded Excerpt

Excerpt from first movement of Trio Sonata in F Minor, by Sammartini.

Historical Classification, Item No. 17

Question

Is this excerpt most representative of the Baroque, Classical, Romantic, or Modern Period?

Answer Array

B C R M

Notational Display

(none)

Contents of Recorded Excerpt

Excerpt from second movement of Symphony No. 1, by Mahler.

APPENDIX B

SUMMARY TABLES FOR MEDLEY PROCEDURE DATA

Table 16

Medley Procedure for Omitted Notes, All Students

Source of Variation	DF	SS	MS
A. Group receiving computerized version			
(1) Students	31	29.9984	0.9677
(2) Items	19	57.4984	3.0262
(3) Error	589	62.2516	0.1057
(4) Total	639	149.7484	
B. Group receiving conventional version			
(5) Students	27	16.8214	0.6230
(6) Items	19	27.0785	1.4252
(7) Error	513	80.3215	0.1566
(8) Total	559	124.2214	
C. Combined groups			
(9) Students	59	47.3292	0.8022
(10) Items	19	77.9292	4.1015
(11) Error	1121	149.2208	0.1331
(12) Total	1199	274.4792	
D. Analysis of equivalence			
(13) Groups (9-1-5)	1	0.5094	0.5094
(14) Students (1 + 5)	58	46.8198	0.8072
(15) Items (10)	19	77.9292	4.1015
(16) Error between versions (11-3-7)	19	6.6477	0.3499
(17) Error within versions (3 + 7)	1102	142.5731	0.1294
(18) Total (12)	1199	274.4792	

Test for Criterion 1: Are students ranked in same order?

$\underline{F} = \frac{MS(16)}{MS(17)} = 2.7040$; $p < .005$, criterion not met.

Test for Criterion 2: Are variances of errors of measurement equal?

$\underline{F} = \frac{MS(7)}{MS(3)} = 1.4816$; $p < .005$, criterion not met.

Test for Criterion 3: Are variances of obtained scores equal?

$\underline{F} = \frac{MS(1)}{MS(5)} = 1.5533$; NS, criterion met.

Test for Criterion 4: Are means equal?

$\underline{F} = \frac{MS(14)}{MS(13)} = 1.5846$; NS, criterion met.

Table 17

Medley Procedure for Omitted Notes,
Lower-term Students Only

Source of Variation	DF	SS	MS
A. Group receiving computerized version			
(1) Students	17	17.9250	1.0544
(2) Items	19	39.0972	2.0577
(3) Error	323	32.3528	0.1002
(4) Total	359	89.3750	
B. Group receiving conventional version			
(5) Students	15	10.7875	0.7192
(6) Items	19	14.3875	0.7572
(7) Error	285	45.7125	0.1604
(8) Total	319	70.8875	
C. Combined groups			
(9) Students	33	31.4485	0.9530
(10) Items	19	47.0867	2.4782
(11) Error	627	84.4633	0.1347
(12) Total	679	162.9985	
D. Analysis of equivalence			
(13) Groups (9-1-5)	1	2.7360	2.7360
(14) Students (1 + 5)	32	28.7125	0.8973
(15) Items (10)	19	47.0867	2.4782
(16) Error between versions (11-3-7)	19	6.3980	0.3367
(17) Error within versions (3 + 7)	608	78.0653	0.1284
(18) Total (12)	679	162.9985	

Test for Criterion 1: Are students ranked in same order?

$\underline{F} = \frac{MS(16)}{MS(17)}$ = 2.6223; p <.005, criterion not met.

Test for Criterion 2: Are variances of errors of measurement equal?

$\underline{F} = \frac{MS(7)}{MS(3)}$ = 1.6008; p <.001, criterion not met.

Test for Criterion 3: Are variances of obtained scores equal?

$\underline{F} = \frac{MS(1)}{MS(5)}$ = 1.4661; NS, criterion met.

Test for Criterion 4: Are means equal?

$\underline{F} = \frac{MS(13)}{MS(14)}$ = 3.0491; NS, criterion met.

Table 18

Medley Procedure for Omitted Notes,
Upper-term Students Only

Source of Variation	DF	SS	MS
A. Group receiving computerized version			
(1) Students	13	6.1429	0.4725
(2) Items	19	23.4429	1.2338
(3) Error	247	24.8571	0.1006
(4) Total	279	54.4429	
B. Group receiving conventional version			
(5) Students	11	6.0333	0.5485
(6) Items	19	14.6666	0.7719
(7) Error	209	35.9668	0.1721
(8) Total	239	56.6667	
C. Combined groups			
(9) Students	25	12.7923	0.5117
(10) Items	19	34.2385	1.8020
(11) Error	475	61.3615	0.1292
(12) Total	519	108.3923	
D. Analysis of equivalence			
(13) Groups (9-1-5)	1	0.6161	0.6161
(14) Students (1 + 5)	24	12.1762	0.5073
(15) Items (10)	19	34.2385	1.8020
(16) Error between versions (11-3-7)	19	0.5376	0.0283
(17) Error within versions (3 + 7)	456	60.8239	0.1334
(18) Total (12)	519	108.3923	

Test for Criterion 1: Are students ranked in same order?

$F = \frac{MS_{(17)}}{MS_{(16)}} = 4.7138$; $p < .001$, criterion not met.

Test for Criterion 2: Are variances of errors of measurement equal?

$F = \frac{MS_{(7)}}{MS_{(3)}} = 1.7107$; $p < .001$, criterion not met.

Test for Criterion 3: Are variances of obtained scores equal?

$F = \frac{MS_{(5)}}{MS_{(1)}} = 1.1608$; NS, criterion met.

Test for Criterion 4: Are means equal?

$F = \frac{MS_{(13)}}{MS_{(14)}} = 1.2145$; NS, criterion met.

Table 19

Medley Procedure for Overall Rhythmic Inaccuracies, All Students

Source of Variation	DF	SS	MS
A. Group receiving computerized version			
(1) Students	31	32.7609	1.0568
(2) Items	19	62.0171	3.2641
(3) Error	589	60.8329	0.1033
(4) Total	639	155.6109	
B. Group receiving conventional version			
(5) Students	27	10.0339	0.3716
(6) Items	19	31.3768	1.6514
(7) Error	513	93.5732	0.1824
(8) Total	559	134.9839	
C. Combined groups			
(9) Students	59	43.2367	0.7328
(10) Items	19	86.8367	4.5702
(11) Error	1121	160.5633	0.1432
(12) Total	1199	290.6367	
D. Analysis of equivalence			
(13) Groups (9-1-5)	1	0.4419	0.4419
(14) Students (1 + 5)	58	42.7948	0.7378
(15) Items (10)	19	86.8367	4.5702
(16) Error between versions (11-3-7)	19	6.1572	0.3241
(17) Error within versions (3 + 7)	1102	154.4061	0.1401
(18) Total (12)	1199	290.6367	

Test for Criterion 1: Are students ranked in same order?

$\underline{F} = \frac{MS_{(16)}}{MS_{(17)}} = 2.3133$; $p < .001$, criterion not met.

Test for Criterion 2: Are variances of errors of measurement equal?

$\underline{F} = \frac{MS_{(7)}}{MS_{(3)}} = 1.7657$; $p < .001$, criterion not met.

Test for Criterion 3: Are variances of obtained scores equal?

$\underline{F} = \frac{MS_{(1)}}{MS_{(5)}} = 2.8439$; $p < .005$, criterion not met.

Test for Criterion 4: Are means equal?

$\underline{F} = \frac{MS_{(14)}}{MS_{(13)}} = 1.6696$; NS, criterion met.

Table 20

Medley Procedure for Overall Rhythmic Inaccuracies, Lower-term Students Only

Source of Variation	DF	SS	MS
A. Group receiving computerized version			
(1) Students	17	20.7000	1.2176
(2) Items	19	34.3222	1.8064
(3) Error	323	34.0778	0.1055
(4) Total	359	89.1000	
B. Group receiving conventional version			
(5) Students	15	7.3719	0.4915
(6) Items	19	16.4344	0.8650
(7) Error	285	54.8156	0.1923
(8) Total	319	78.6219	
C. Combined groups			
(9) Students	33	28.1132	0.8519
(10) Items	19	45.7338	2.4070
(11) Error	627	93.9162	0.1498
(12) Total	679	167.7632	
D. Analysis of equivalence			
(13) Groups (9-1-5)	1	0.0413	0.0413
(14) Students (1 + 5)	32	28.0719	0.8772
(15) Items (10)	19	45.7338	2.4070
(15) Error between versions (11-3-7)	19	5.0228	0.2644
(17) Error within versions (3 + 7)	608	88.8934	0.1462
(18) Total (12)	679	167.7632	

Test for Criterion 1: Are students ranked in same order?

$\underline{F} = \frac{MS_{(16)}}{MS_{(17)}}$ = 1.8085; p <.05, criterion not met.

Test for Criterion 2: Are variances of errors of measurement equal?

$\underline{F} = \frac{MS_{(7)}}{MS_{(3)}}$ = 1.8227; p <.001, criterion not met.

Test for Criterion 3: Are variances of obtained scores equal?

$\underline{F} = \frac{MS_{(1)}}{MS_{(5)}}$ = 2.4773; p <.05, criterion not met.

Test for Criterion 4: Are means equal?

$\underline{F} = \frac{MS_{(14)}}{MS_{(13)}}$ = 21.2397; NS, criterion met.

Table 21

Medley Procedure for Overall Rhythmic Inaccuracies,
Upper-term Students Only

Source of Variation	DF	SS	MS
A. Group receiving computerized version			
(1) Students	13	11.5750	0.8904
(2) Items	19	29.4107	1.5479
(3) Error	247	24.6393	0.9975
(4) Total	279	65.6250	
B. Group receiving conventional version			
(5) Students	11	2.0333	0.1848
(6) Items	19	16.7333	0.8807
(7) Error	209	36.9667	0.1769
(8) Total	239	55.7333	
C. Combined groups			
(9) Students	25	13.6173	0.5447
(10) Items	19	42.7904	2.2521
(11) Error	475	64.9596	0.1368
(12) Total	519	121.3673	
D. Analysis of equivalence			
(13) Groups (9-1-5)	1	0.0090	0.0090
(14) Students (1 + 5)	24	13.6083	0.5670
(15) Items (10)	19	42.7904	2.2521
(16) Error between versions (11-3-7)	19	3.3536	0.1765
(17) Error within versions (3 + 7)	456	61.6060	0.1351
(18) Total (12)	519	121.3673	

Test for Criterion 1: Are students ranked in same order?

$\underline{F} = \frac{MS_{(16)}}{MS_{(17)}}$ = 1.3064; NS, criterion met.

Test for Criterion 2: Are variances of errors of measurement equal?

$\underline{F} = \frac{MS_{(3)}}{MS_{(7)}}$ = 5.6388, $p < .001$, criterion not met.

Test for Criterion 3: Are variances of obtained scores equal?

$\underline{F} = \frac{MS_{(1)}}{MS_{(5)}}$ = 4.8182; $p < .01$, criterion not met.

Test for Criterion 4: Are means equal?

$\underline{F} = \frac{MS_{(14)}}{MS_{(13)}}$ = 63.0000; NS, criterion met.

Table 22

Medley Procedure for Faulty Interpretations,
All Students

Source of Variation	DF	SS	MS
A. Group receiving computerized version			
(1) Students	31	30.6750	0.9895
(2) Items	19	61.8750	3.2566
(3) Error	589	61.8250	0.1050
(4) Total	639	154.3750	
B. Group receiving conventional version			
(5) Students	27	9.1214	0.3378
(6) Items	19	20.6500	1.0868
(7) Error	513	100.4500	0.1958
(8) Total	559	130.2214	
C. Combined groups			
(9) Students	59	40.2367	0.6820
(10) Items	19	72.3367	3.8072
(11) Error	1121	172.4633	0.1538
(12) Total	1191	285.0367	
D. Analysis of equivalence			
(13) Groups (9-1-5)	1	0.4403	0.4403
(14) Students (1 + 5)	58	39.7964	0.6861
(15) Items (10)	19	72.3367	3.8072
(16) Error between versions (11-3-7)	19	10.1883	0.5362
(17) Error within versions (3 + 7)	1102	162.2750	0.1473
(18) Total (12)	1199	285.0367	

Test for Criterion 1: Are students ranked in same order?

$\underline{F} = \frac{MS_{(16)}}{MS_{(17)}}$ = 3.6402; p <.001, criterion not met.

Test for Criterion 2: Are variances of errors of measurement equal?

$\underline{F} = \frac{MS_{(7)}}{MS_{(3)}}$ = 1.8648; p <.001, criterion not met.

Test for Criterion 3: Are variances of errors of obtained scores equal?

$\underline{F} = \frac{MS_{(1)}}{MS_{(5)}}$ = 2.9292; p <.001, criterion not met.

Test for Criterion 4: Are means equal?

$\underline{F} = \frac{MS_{(14)}}{MS_{(13)}}$ = 1.5583, NS, criterion met.

Table 23

Medley Procedure for Faulty Interpretations, Lower-term Students Only

Source of Variation	DF	SS	MS
A. Group receiving computerized version			
(1) Students	17	18.7274	1.1028
(2) Items	19	34.9194	1.8379
(3) Error	323	35.5306	0.1100
(4) Total	359	89.1972	
B. Group receiving conventional version			
(5) Student	15	6.4969	0.4331
(6) Items	19	12.6844	0.6676
(7) Error	285	56.0656	0.1967
(8) Total	319	75.2469	
C. Combined groups			
(9) Students	33	26.1882	0.7936
(10) Items	19	41.0353	2.1598
(11) Error	627	98.1647	0.1566
(12) Total	679	165.3882	
D. Analysis of equivalence			
(13) Groups (9-1-5)	1	0.9441	0.9441
(14) Students (1 + 5)	32	25.2441	0.7889
(15) Items (10)	19	41.0353	2.1598
(16) Error between versions (11-3-7)	19	6.5685	0.3457
(17) Error within versions (3 + 7)	608	91.5962	0.1507
(18) Total (12)	679	165.3882	

Test for Criterion 1: Are students ranked in same order?

$\underline{F} = \frac{MS_{(16)}}{MS_{(17)}} = 2.2940$; $p < .001$, criterion not met.

Test for Criterion 2: Are variances of errors of measurement equal?

$\underline{F} = \frac{MS_{(7)}}{MS_{(3)}} = 1.7882$; $p < .001$, criterion not met.

Test for Criterion 3: Are variances of obtained scores equal?

$\underline{F} = \frac{MS_{(1)}}{MS_{(5)}} = 2.5463$; $p < .05$, criterion not met.

Test for Criterion 4: Are means equal?

$\underline{F} = \frac{MS_{(13)}}{MS_{(14)}} = 1.1967$; NS, criterion met.

Table 24

Medley Procedure for Faulty Interpretations, Upper-term Students Only

Source of Variation	DF	SS	MS
A. Group receiving computerized version			
(1) Students	13	10.1464	0.7805
(2) Items	19	28.8964	1.5209
(3) Error	247	24.3536	0.9860
(4) Total	279	63.3964	
B. Group receiving conventional version			
(5) Students	11	2.5458	0.2314
(6) Items	19	12.4791	0.6568
(7) Errors	209	39.8709	0.1908
(8) Total	239	54.8958	
C. Combined groups			
(9) Students	25	12.7000	0.5080
(10) Items	19	34.3000	1.8053
(11) Error	475	71.3000	0.1501
(12) Total	519	118.3000	
D. Analysis of equivalence			
(13) Groups (9-1-5)	1	0.0078	0.0078
(14) Students (1 + 5)	24	12.6922	0.5288
(15) Items (10)	19	34.3000	1.8053
(16) Error between versions (11-3-7)	19	7.0755	0.3724
(17) Error within versions (3 + 7)	456	64.2245	0.1408
(18) Total (12)	519	118.3000	

Test for Criterion 1: Are students ranked in same order?

$F = \frac{MS(16)}{MS(17)} = 2.6449$; $p < .001$, criterion not met.

Test for Criterion 2: Are variances of errors of measurement equal?

$F = \frac{MS(3)}{MS(7)} = 5.1677$; $p < .001$, criterion not met.

Test for Criterion 3: Are variances of obtained scores equal?

$F = \frac{MS(1)}{MS(5)} = 3.3729$; $p < .05$, criterion not met.

Test for Criterion 4: Are means equal?

$F = \frac{MS(14)}{MS(13)} = 67.7949$; NS, criterion met.

Table 25

Medley Procedure for Historical Classification, All Students

Source of Variation	DF	SS	MS
A. Group receiving computerized version			
(1) Students	31	32.7938	1.0579
(2) Items	19	52.2188	2.7484
(3) Error	589	71.0812	0.1207
(4) Total	639	156.0938	
B. Group receiving conventional version			
(5) Students	27	10.0054	0.3706
(6) Items	19	26.2697	1.3826
(7) Error	513	101.2803	0.1974
(8) Total	559	137.5554	
C. Combined groups			
(9) Students	59	49.0092	0.8307
(10) Items	19	75.9425	3.9970
(11) Error	1121	174.9075	0.1560
(12) Total	1199	299.8592	
D. Analysis of equivalence			
(13) Groups (9-1-5)	1	6.2100	6.2100
(14) Students (1 + 5)	58	42.7992	0.7379
(15) Items (10)	19	75.9425	3.9970
(16) Error between versions (11-3-7)	19	2.5460	0.1340
(17) Error within versions (3 + 7)	1102	172.3615	0.1564
(18) Total (12)	1199	299.8592	

Test for Criterion 1: Are students ranked in same order?

$F = \frac{MS_{(17)}}{MS_{(16)}}$ + 1.1672; NS, criterion met.

Test for Criterion 2: Are variances of errors of measurement equal?

$F = \frac{MS_{(7)}}{MS_{(3)}}$ = 1.6355; p <.001, criterion not met.

Test for Criterion 3: Are variances of obtained scores equal?

$F = \frac{MS_{(1)}}{MS_{(5)}}$ = 2.8546; p <.005, criterion not met.

Test for Criterion 4: Are means equal?

$F = \frac{MS_{(13)}}{MS_{(14)}}$ = 8.4158; p <.01, criterion not met.

Table 26

Medley Procedure for Historical Classification,
Lower-term Students Only

Source of Variation	DF	SS	MS
A. Group receiving computerized version			
(1) Students	17	13.5250	0.7956
(2) Items	19	36.7194	1.9326
(3) Error	323	32.5306	0.1007
(4) Total	359	82.7750	
B. Group receiving conventional version			
(5) Students	15	3.6875	0.2458
(6) Item	19	17.9375	0.9441
(7) Error	285	58.0625	0.2037
(8) Total	319	79.6875	
C. Combined groups			
(9) Students	33	22.2779	0.6751
(10) Items	19	44.9103	2.3637
(11) Error	627	100.3397	0.1600
(12) Total	679	167.5279	
D. Analysis of equivalence			
(13) Groups (9-1-5)	1	5.0654	5.0654
(14) Students (1 + 5)	32	17.2125	0.5379
(15) Items (10)	19	44.9103	2.3637
(16) Error between versions (11-3-7)	19	9.7466	0.5130
(17) Error within versions (3 + 7)	608	90.5931	0.1490
(18) Total (12)	679	167.5279	

Test for Criterion 1: Are students ranked in same order?

$\underline{F} = \frac{MS_{(16)}}{MS_{(17)}} = 3.4430$; $p < .001$, criterion not met.

Test for Criterion 2: Are variances of errors of measurement equal?

$\underline{F} = \frac{MS_{(7)}}{MS_{(3)}} = 2.0228$; $p < .001$, criterion not met.

Test for Criterion 3: Are variances of obtained scores equal?

$\underline{F} = \frac{MS_{(1)}}{MS_{(5)}} = 3.2368$; $p < .025$, criterion not met.

Test for Criterion 4: Are means equal?

$\underline{F} = \frac{MS_{(13)}}{MS_{(14)}} = 9.4170$; $p < .005$, criterion not met.

Table 27

Medley Procedure for Historical Classification, Upper-term Students Only

Source of Variation	DF	SS	MS
A. Group receiving computerized version			
(1) Students	13	15.9464	1.2266
(2) Items	19	19.6393	1.0336
(3) Error	247	34.4107	0.1393
(4) Total	279	69.9964	
B. Group receiving conventional version			
(5) Students	11	5.4125	0.4920
(6) Items	19	14.8792	0.7831
(7) Error	209	36.6708	0.1755
(8) Total	238	56.9625	
C. Combined groups			
(9) Students	25	22.8923	0.9157
(10) Items	19	28.5692	1.5036
(11) Error	475	77.0308	0.1622
(12) Total	519	128.4923	
D. Analysis of equivalence			
(13) Groups (9-1-5)	1	1.5334	1.5334
(14) Students (1 + 5)	24	21.3589	0.8900
(15) Items (10)	19	28.5692	1.5036
(16) Error between versions (11-3-7)	19	5.9493	0.3131
(17) Error within versions (3 + 7)	456	71.0815	0.1559
(18) Total (12)	519	128.4923	

Test for Criterion 1: Are students ranked in same order?

$\underline{F} = \frac{MS_{(16)}}{MS_{(17)}} = 2.0083$; $p < .01$, criterion not met.

Test for Criterion 2: Are variances of errors of measurement equal?

$\underline{F} = \frac{MS_{(7)}}{MS_{(3)}} = 1.2599$; $p < .05$, criterion not met.

Test for Criterion 3: Are variances of obtained scores equal?

$\underline{F} = \frac{MS_{(1)}}{MS_{(5)}} = 2.4931$, NS, criterion met.

Test for Criterion 4: Are means equal?

$\underline{F} = \frac{MS_{(13)}}{MS_{(14)}} = 1.7229$; NS, criterion met.

APPENDIX C

QUESTIONNAIRE ITEMS APPENDED TO BOTH TEST VERSIONS

QUESTIONNAIRE ITEMS APPENDED TO BOTH TEST VERSIONS

1. Of the four sections, I thought that the most difficult section for me was the

    ____(A) Omitted Notes section

    ____(B) Overall Rhythmic Inaccuracies section

    ____(C) Faulty Interpretation section

    ____(D) Historical Classification section

2. Of the four sections, I thought that the least difficult section for me was the

    ____(A) Omitted Notes section

    ____(B) Overall Rhythmic Inaccuracies section

    ____(C) Faulty Interpretation section

    ____(D) Historical Classification section

3. The overall quality of sound reproduction was generally

    ____(A) very poor and distracting; it made the questions difficult to answer.

    ____(B) not good, but it did not interfere with my ability to answer the questions.

    ____(C) fair; it certainly was adequate for the test.

    ____(D) quite good; it was often enjoyable to listen.

4. The overall quality of the notation was generally

    ____(A) very poor; the illegibility of the notes often made it difficult to answer questions.

    ____(B) not good; but it did not interfere with my ability to answer the questions.

    ____(C) not comparable to printed music, but it was certainly adequate for the purpose.

    ____(D) quite good; it was comparable to printed music in most respects.

5. With regard to the speed of the test, I think that the test generally moved

____(A) too slowly; there was needless delay between items.

____(B) at a comfortable pace for me.

____(C) too rapidly; there was insufficient time between items.

6. While I was taking the test, I generally felt

____(A) quite calm and relaxed; there was very little pressure on me.

____(B) slightly tense; there was some pressure on me, but it was largely of my own making.

____(C) rather tense; pressure was being placed upon me by the testing situation.

____(D) quite tense and agitated; I was constantly being pressured and urged to produce answers.

7. If I had a choice, I would have preferred to take the test in

____(A) the CAI Laboratory, using the computerized instructional station in an individualized manner.

____(B) a conventional paper-and-pencil testing situation, as part of a group taking the test simultaneously.

____(C) It really made no difference.

APPENDIX D

EXAMPLE OF COURSEWRITER PROGRAMMING

EXAMPLE OF COURSEWRITER PROGRAMMING

The example below illustrates use of the Coursewriter II computer language. Literacy in Coursewriter is required to interpret the statements, but, esentially, the computer is told what alphameric characters to display and where on the cathode ray tube screen to display them, what student responses to expect and what action is to be taken for each response, what scores to store and where to store them, when to play a tape segment or display an image, when to query a student, and how long to allow for his response. The example includes the programming for the ninth, tenth, eleventh, and twelfth FI items.

```
FI9*E
     1  PR  *E
     2  DE  0+/32*E
     3  FPI 99*E
     4  DT  0,5+/4,0+/40,0+/(W)HAT IS WRONG WITH THE PERFORMER'S*C*I
            INTERPRETATION OF THIS MELODY(/*E
     5  DT  7,5+/2,7+/35,5+/+, (T)HE ARTICULATION IS INCORRECT.*E
     6  DT  13,5+/2,13+/35,5+/+, (T)HE DYNAMICS ARE UNOBSERVED.*E
     7  DT  19,5+/2,19+/35,5+/+, (T)HE TEMPO IS INAPPROPRIATE.*E
     8  DT  25,5+/2,25+/35,5+/+, (T)HE RUBATO IS INSUFFICIENT.*E
     9  PA  70*E
    10  AUP FIO*E1040,0+/48*E
    11  EPP 450+/QUFI9*E
    12  NX  *E
    13  BR  PR1*E
    14  CAP 4,12,3,4+/cc*E
    15  SB  C6+/C6*E
    16  AD  1+/C3*E
    17  BR  PR2*E
    18  WAP 4,6,3,4+/W1*E
    19  WBP 4,18,3,4+/W3*E
    20  WBP 4,24,3,4+/W4*E
    21  BR  PR1*E
    22  UN  UU*E
    23  DT  28,7+/2,28+/33,7+/(T)OUCH ONLY A +,.*E
    24  PA  40*E
    25  DE  28+/2*E
    26  PR  *E
    27  AD  1+/C5*E
    28  AD  1+/C6*E
    29  BR   HCTR1+/C6+/E+/3*E
    30  BR   HCTR1+/C5+/E+/5*E
FI10*E
     1  PR   *E
     2  DE   0+/32*E
     3  FP1  100*E
     4  DT   0,5+/4,0+/40,0+/(W)HAT IS WRONG WITH THE PERFORMER'S*C*I
             INTERPRETATION OF THIS MELODY(/*E
     5  DT   7,5+/2,7+/35,5+/+, (T)HE TEMPO IS INAPPROPRIATE.*E
     6  DT   13,5+/2,13+/35,5+/+, (T)HE RUBATO IS INSUFFICIENT.*E
     7  DT   19,5+/2,19+/35,5+/+, (T)HE RUBATO IS EXCESSIVE.*E
     8  DT   25,5+/2,25+/35,5+/+, (T)HE DYNAMICS ARE UNOBSERVED.*E
     9  PA   70*E
    10  AUP  FI10*E1089,0+/64*E
    11  EPP  450+/QUFI10*E
    12  NX   *E
    13  BR   PR1*E
    14  CAP  4,12,3,4+/CC*E
    15  SB   C6+/C6*E
    16  AD   1+/C3*E
    17  BR   PR2*E
    18  WAP  4,6,3,4+/W1*E
    19  WBP  4,18,3,4+/W3*E
    20  WBP  4,24,3,4+/W4*E
    21  BR   PR1*E
    22  UN   UU*E
    23  DT   28,7+/2,28+/33,7+/(T)OUCH ONLY A +,.*E
    24  PA   40*E
    25  DE   28+/2*E
    26  PR   *E
    27  AD   1+/C5*E
    28  AD   1+/C6*E
    29  BR   HCTR1+/C6+/E+/3*E
    30  BR   HCTR1+/c5+/E+/5*E
FI11*E
     1  PR   *E
     2  DE   0+/32*E
     3  FP1  111*E
     4  DT   0,5+/4,0+/40,0+.(W)HAT IS WRONG WITH THE PERFORMER'S*C*I
             INTERPRETATION OF THIS MELODY(/*E
     5  DT   7,5+/2,7+/35,3+/+, (T)HE TEMPO IS INAPPROPRIATE.*E
     6  DT   13,5+/2,13+/35,5+/+, (T)HE ARTICULATION IS INCORRECT.*E
     7  DT   19,5+/2,19+/35,5+/+, (T)HE DYNAMICS ARE UNOBSERVED.*E
     8  DT   25,5+/2,25+/35,5+/+, (T)HE RUBATO IS EXCESSIVE.*E
     9  PA   70*E
    10  AUP  FI11*E1098,1+/94*E
    11  EPP  450+/QUFI11*E
    12  NX   *E
    13  BR   PR1*E
    14  CAP  4,12,3,4+/CC*E
    15  SB   C6+/C6*E
    16  AD   1+/C3*E
    17  BE   PR2*E
```

```
      18  WAP 4,6,3,4+/W1*E
      19  WBP 4,18,3,4+/W3*E
      20  WBP 4,24,3,4+/W4*E
      21  BR  PR1*E
      22  UN  UU*E
      23  DT  28,7+/2,28+/33,7+/(T)OUCH ONLY A +,.*E
      24  PA  40*E
      25  DE  28+/2*E
      26  PR  *E
      27  AD  1+/C5*E
      28  AD  1+/C6*E
      29  BR  HCTR1+/C6+/E+/3*E
      30  BR  HCTR1+/C5+/E+/5*E
FI12*E
       1  PR  *E
       2  BR  PR2+/S3+/1*E
       3  LD  1+/S3*E
       4  DE  0+/32*E
       5  FP1 112*E
       6  DT  0,5+/4,0+/40,0+/(W)HAT IS WRONG WITH THE PERFORMER'S*C*I
              INTERPRETATION OF THIS MELODY(/*E
       7  DT  7,5+/2,7+/35,5+/+, (T)HE TEMPO IS INAPPROPRIATE.*E
       8  DT  13,5+/2,13+/35,5+/+, (T)HE RUBATO IS EXCESSIVE.*E
       9  DT  19,5+/2,19+/35,5+/+, (T)HE DYNAMICS ARE UNOBSERVED.*E
      10  DT  25,5+/2,25+/35,5+/+, (T)HE ARTICULATION IS INCORRECT.*E
      11  PA  70*E
      12  AUP FI12*E1110,2+/86*E
      13  EPP 450+/QUF112*E
      14  NX  *E
      15  BR  PR1*E
      16  CAP 4,6,3,4+/CC*E
      17  SB  C6+/C6*E
      18  AD  1+/C3*E
      19  BR  PR2+/c5+/G+/0*E
      20  AD  3+/C3*E
      21  BR  FI16*E
      22  WAP 4,12,3,4+/W2*E
      23  WBP 4,18,3,4+/W3*E
      24  WBP 4,24,3,4+/W4*E
      25  BR  PR1*E
      26  UN  UU*E
      27  DT  28,7+/2,28+/33,7+/(T)OUCH ONLY A +,.*E
      28  PA  40*E
      29  DE  28+/2*E
      30  PR  *E
      31  AD  1+/C5*E
      32  AD  1+/C6*E
      33  BR  HCTR1+/C6+/E+/3*E
      34  BR  HCTR1+/C5+/E+/5*E
      35  BR  FI9+/C5+/E+/1*E
```

APPENDIX E

EXAMPLE OF STUDENT RECORDS

EXAMPLE OF STUDENT RECORDS

Detailed information regarding any student's performance on the IBM 1500 Instructional System is available through student records. This example lists the performance records for six students on the fourth item of the ORI scale. Information contained includes the code number of the question (QUORI4 in this example), the code numbers of the students, time elapsed between the end of the playing of the taped musical example and the students' responses, the response code and location of the students' response, and the dates and times of the responses.

```
COURSE   SEG   S      EP IDENT.   LATENCY   MATCH   DATE       TIME
MUTES     0    X27    QUORI4       39.5      CC     10/8/70    14:53.93
                         RESPONSE - ROW 24 COL C5
MUTES     0    X28    QUORI4       12.1      W1     10/6/70    13:38.7
                         RESPONSE - ROW 06 COL C5
MUTES     0    X29    QUORI4       11.3      CC     10/5/70    14:51.1
                         RESPONSE - ROW 24 COL C5
MUTES     0    X30    QUORI4        4.0      CC     10/9/70    10:31.30
                         RESPONSE - ROW 25 COL C5
MUTES     0    X31    QUORI4       13.3      CC     10/5/70    14:42.39
                         RESPONSE - ROW 24 COL C5
MUTES     0    X33    QUORI4       12.5      W1     10/8/70     9:57.89
                         RESPONSE - ROW 06 COL C5
```

APPENDIX F

SELECTION OF TESTS FOR PROGRAMMING

SELECTION OF TESTS FOR PROGRAMMING

There were five tests from which to select in the pitch area. The EN test was administered to a group of students too small for the establishment of meaningful descriptive statistics. (See Chapter III, pages 44-45.) The item difficulty indices tend to be weighted toward the less difficult end of the MI scale (Table 1, page 63); the TC test may have contained, at the time of empirical trials, unrealistically difficult items (Chapter II, page 44). The HI test after trial was found to contain a sparsity of easy ($p \geq .70$) items (Chapter III, page 43). The ON test does not contain the problems associated with the other pitch tests; furthermore, the melodies of the ON test are of greater musical interest than isolated triads and intervals. Therefore, the ON test was selected for programming.

Three tests were developed in the rhythm area; the strictly visual IMS test was never administered (Chapter III, page 48). The ORI test was selected for programming in preference to the RD test because the ORI item difficulty indices are spaced at more nearly equal intervals than the RD item difficulty indices (Table 1, page 63), and the raw data obtained from the trial of the RD test might have been confounded by student response patterns (Chapter III, page 45).

In the interpretation area, there were two tests from which to select. The BP and FI tests were each successfully administered; the descriptive statistics obtained were similar (Table 2, page 66). However, each test was administered at two institutions, and the rank-order correlation between difficulty scales obtained at the respective pairs of institutions favored the FI test (Table 3, page 70). Furthermore, the FI test, albeit subjective (Chapter III, pages 37-38), requires, in the opinion of the researcher, a broader range of thinking than the identification of unmusical interruptions in the BP test, and the FI test was selected for programming.

The HC(L) version of the HC test was selected for programming in the style area primarily because it was the one test that had been administered at The Pennsylvania State University. The HC(Y) version evidently is dissimilar to the HC(L) version because the rank-order correlation of the difficulty rankings for each test is not close to 1.00 (Chapter III, page 50). It was not possible to administer both HC versions to the same students, hence the decision was made to program the HC(L) version.

OE 6000 (REV. 9-66)

DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
OFFICE OF EDUCATION

ERIC REPORT RESUME

ERIC ACCESSION NO.

CLEARINGHOUSE ACCESSION NUMBER

RESUME DATE — —

P.A.

T.A.

IS DOCUMENT COPYRIGHTED? YES ☐ NO ☐

ERIC REPRODUCTION RELEASE? YES ☐ NO ☐

TITLE

Development of a Test for the Nonperformance Aspects of Music Education

PERSONAL AUTHOR(S)

Radocy, Rudolf E.

INSTITUTION (SOURCE)

The Pennsylvania State University, University Park, Pa. 16802

SOURCE CODE

REPORT/SERIES NO.

OTHER SOURCE

Computer Assisted Instruction Laboratory, 201 Chambers Building

SOURCE CODE R-38

OTHER REPORT NO.

OTHER SOURCE

SOURCE CODE

OTHER REPORT NO.

PUBL. DATE — —

CONTRACT/GRANT NUMBER OEG-2-700018(509)

PAGINATION, ETC.

RETRIEVAL TERMS

computerized testing; criterion-referenced testing; incremental testing; music education; music testing; sequential testing; nonperformance musical behaviors

IDENTIFIERS

IBM 1500 Instructional System

ABSTRACT

A prototype computerized criterion-referenced test of certain nonperformance musical behaviors was developed at The Pennsylvania State University, with the expectation that the test could provide a pattern for development in similar situations.

A total of 783 criterion-referenced test items were administered to undergraduates. Item difficulty indices were computed, and twenty-item scales, arranged in order of difficulty, were selected for each of twelve subtests. Four subtests were programmed for the IBM 1500 Instructional System.

A sequential or incremental programming strategy was adopted. A student receives every fourth item of each twenty-item scale until he makes an initial error. A reverse branch of three then occurs; the forward increment is changed to one. Each subtest is terminated when three successive errors occur, a total of five errors occurs, or the end of the scale is reached. Nonadministered items are assumed to be correct if they are of less difficulty than the most difficult correctly answered administered item.

While not statistically equivalent to an off-line version of the same test, the computerized test performs adequately from a qualitative standpoint. Refinement by reordering of the items on the basis of more stable indices of difficulty is recommended for quantitative improvement.